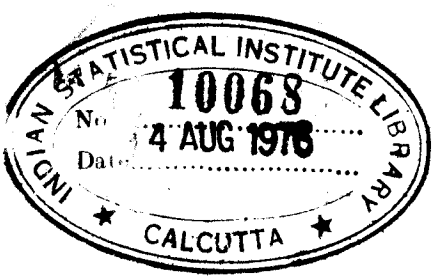
SOME STATISTICAL STUDIES ON LANGUAGES

N. Bhattacharya



Indian Statistical Institute,
Calcutta-35.
October 1965.

'IST'

Preface and Acknowledgements

The investigations reported in the present thesis were started towards the end of 1957 and carried along intermittently amidst teaching and professional work on sample surveys and econometrics. Curiosity provided the major impulse. The author's ambition was, initially, to throw up some numerical tables for Bengali, Sanskrit, Prakrit and Pali, such as are found for many western languages in Herdan's (1956) Language as Choice and Chance. Gradually, as the work progressed, the view changed and more analytical studies suggested themselves. Some of these latter have been completed in a provisional manner and reported in the present thesis.

Chapter 1 attempts a broad survey of previous researches classified under three heads : (i) statistical studies on literary style, based on word-length, sentence-length, size and composition of vocabulary etc. (ii) studies on statistical properties of languages, e.g, the relative frequencies of letters or phonemes, and the Zipf law of word-frequencies, and (iii) information-theoretic analyses of languages carried out by Shannon and others. One section is devoted to Indian work in various lines.

Chapters 2 to 7 report on studies relating to word-length, almost entirely confined to Bengali, with emphasis on prose fiction. Word-length has been measured in syllables. Chapters 8 and 9 describe the corresponding investigations on sentence-length, measuring sentence-length in terms of the number of words; but the scale of investigation is more modest here and poetry has been excluded for obvious reasons.

Chapter 2 gives an account of the samples of Bengali words analysed in the different studies. Probability samples of words were selected from many prose works, but as the study progressed, it became apparent that non-probabilistic systematic samples could be conveniently used as approximations to probability samples. Chapter 2 describes these methods of sampling, establishes the "validity" of the systematic samples and examines the sampling properties of the estimates thrown up by the two types of samples.

Probability sampling has seldom been used for statistical studies on languages; and the uses of non-probabilistic samples have hardly been rigorously justified. Yet statistical methods valid for strictly random samples have been used in a few cases, without due reserve. A major objective of the present study is to use probability sampling for investigations on word-length, sentence-length etc. and also to justify the non-probabilistic systematic samples as approximations to probability samples. The question of sampling error has been constantly kept in view. Standard errors were usually not calculated in the detailed way. The technique of independent and interpenetrating networks of subsamples (IPNS) introduced by Mahalanobis (1946) was found to be extremely serviceable.

Chapter 3 presents the word-length distributions, averages etc. for different works, including poems in Bengali. This enables one to make a broad survey of word-length in different fields of literature. The data also reveal the historical trend in word-length in Bengali prose. Another

important question is discussed here, viz., how far word-length statistics can indicate an author's style, that is to say, how far word-length distributions tend to be similar for different works by the same author. The answer seems to be largely in the negative.

Chapter 4 distinguishes between "conversational" words and "other" words used in fiction and examines the word-length distributions and averages separately for the two classes of words for each work. Conversational words are shorter, on the average, than other words. The two averages of word-length and the percentage of words falling under "conversation" lead to a clearer understanding of the variation in word-length between works and over time.

Chapter 5 investigates the randomness of some series of word-lengths in Bengali prose. The autocorrelation coefficient r_1 between lengths of consecutive words is generally significant but small, about 0.06, on the average; the coefficients r_2, r_3 etc. are also similar in magnitude. Many texts, however, contain short passages with unusually high or low levels of word-length. The distinction between conversational matter and other matter contributes to this phenomenon and hence to make the autocorrelation positive. Some work is also done on English prose; for two English novels, the coefficient r_1 is about -0.1 or -0.15, word-length being here measured by the number of letters.

Chapter 6 examines the form of the word-length distribution mainly in Bengali prose. Fucks (1955) observed that x-1 follows the Poisson law approximately in eight languages, Arabic proving to be

an exception, [Here x is the word-length in syllables.] But Fucks considered one average distribution for each language. Rigorous examination showed that the fit was not fully satisfactory for all the eight languages, nor for all the individual works in English, German or Russian for which distributions are available. For Bengali works, the Poisson fit is usually poor, but the lognormal curve, tried on English data by Williams (1956) and Herdan (1958), gave much better fit. Two variants of the lognormal hypothesis were tried. For both, the deviations are generally significant but small in the absolute sense.

Chapter 7 introduces a classification of Bengali syllables and studies the relative frequencies of the different types of syllables. Since different types have different lengths, word-length cannot be adequately measured by the total number of syllables. The relative frequencies are fairly stable in Bengali prose, so that word-length comparisons between works can be based on the total number of syllables; but in Bengali poetry the relative frequencies vary considerably depending on theme, mood and meter.

Chapter 8 describes the probability samples of sentences which form the bulk of the material on sentence-length. Sampling properties of the estimates are discussed. Word-length distributions and averages are given for about 20 works in Bengali prose and also for some short stories and essays. This gives dimensional ideas of sentence-length in different types of writing. Here also, appreciable within author differences are found; these throw doubts on the usefulness of sentence-length data as

indicator of individual style.[1] Finally, we test how far the sentence-length distributions are lognormal, and find that the fits are fairly satisfactory for all the works.

In Chapter 9 we apply some tests of randomness on several sentence-length series for Bengali and English. The series are found to be nearly random; the autocorrelation coefficients r_1, r_2, r_3, r_6 and r_{10} are all near zero; r_1 is 0.07, on the average, and significant, but r_{10} is nowhere significant. The validity of some non-probabilistic systematic samples of sentences as approximations to probability samples of sentences is also demonstrated in this chapter.

Chapter 10 analyses some word-counts on written English prose and studies the negative correlation between the frequencies of occurrence and the lengths in letters of words found in these counts. It also demonstrates the economic effect of such negative correlation on the lengths of texts. It was not possible to estimate the correlation or its economic effect in infinitely large counts. For all root words or all particular words, the best available estimates are —0.14 for the correlation coefficient (r) and about 0.3 for the two correlation ratios. The text-length is 62% of the average expected for a random matching of words and meanings. The minimum and maximum values corresponding respectively to most economic and most uneconomic matchings are : —0.20 and +0.28 for r, and 46% and 176% for the text-length as percentage. Similar results are given for content words, function words and nouns.

1/ Generally speaking, one should not be very optimistic about the value of statistical style measures for such purposes; for they may vary with the broad field of literature, with the subject-matter within a given broad field, and also with the age of the author. The literature contains numerous examples of uncritical thinking on these points.

Chapter 11 studies some frequency distributions of lengths of intervals between successive occurrences of the same word. Zipf's (1949) work is re-examined : It is found that for the rare words in "Ulysses", when intervals x are measured in pages, the form $p_x = ax^{-b}$ ($x=1,2,\ldots$; $b \sim 1$) suggested by Zipf does not give a reasonable fit; the geometric distribution is better, but even this is too coarse. The geometric distribution is fairly satisfactory for the interval-distributions for four frequent grammar words ('the', 'to', 'and', 'of') in the English novel, "Pride and Prejudice", and for some fairly frequent grammar words ('se'= he/she, 'sei' = that, etc.) in the Bengali novel, "Gora". Deviations are often significant; generally, short intervals are too frequent compared with the geometric model. Finally, it was shown that the series of intervals are all approximately random. Such studies throw light on the applicability of probability models to texts/speech-samples considered as sequences of words.

Appendix 1 presents some material on word-length in letters in Bengali prose and points to the high correlation between length in syllables and length in letters. Appendix 2 discusses the concentration curves for the discrete distributions of word-length in syllables. Appendix 3 is concerned with word-length in the English novel, "Pride and Prejudice", and reports on studies on the lines followed in Chapters 2,3, 4 and 6. Appendix 4 gives some estimates on the relative frequencies of letters in Bengali prose based on some probability samples of words. Appendix 5 studies the rank-frequency relation for words in Bengali prose, using two counts of very different sizes. Appendix 6 describes

a small-scale experiment which indicates that the time requirements for conveying the same message in writing is more or less the same for English and Bengali.

The author is fully alive to the many limitations of these studies. It is not claimed that the sampling methods actually adopted are optimal in any sense. But a beginning has been made and the way cleared for more conclusive investigations. The sample sizes have been rather/too small in many cases for the large sample assumptions to be safely made. The same applies in several cases to the coverage of works etc. This is, of course, owing to the lack of resources. It should be appreciated that the entire statistical work was done manually without any help from large-scale data processing equipment. It is believed that the conclusions still remain valid or interesting or at least thought-provoking, as the case may be.

Apart from the studies on sampling methods or on the within author variation in word-length or sentence-length, and leaving aside the value of the new material on Bengali or English thrown up for these analyses, the present researches extend the available results in the field in a few new directions. We may mention, for example, the separate analysis of conversational words and other words (Chap.4), the investigations on the randomness of sentence-length series (Chap.9) and interval-length series (Chap.11), and the correlational studies on word-counts contained in Chapter 10. In other cases, previous work was on a very small scale and supplementary evidence had been long overdue; consider, for example, the examination of the lognormality of sentence-length

distributions (Chap. 9). In general, rigorous statistical methods have been used throughout, for testing goodness of fit, randomness of series, etc. etc. where previous workers had depended on general impressions from the data. This alone constitutes an advance whose importance can hardly be over-emphasised.

Professor P. C. Mahalanobis has been a source of inspiration to the author in all these investigations. It was he who initiated linguistic researches in the Indian Statistical Institute many years ago and it was he, again, who drew the author's attention to this particular field of statistical researches. The influence of Professor Mahalanobis on the present researches will be evident to the most casual reader. The author is deeply indebted to him for his kind interest and guidance.

The author is also grateful to Dr. S. K. Chatterjee, now National Professor, for some encouragement in the early stages of the work.

Professor M. Mukherjee of the Indian Statistical Institute made many facilities of work available to the present author; the author is thankful to him and to Professor A. K. Gayen, of the Indian Institute of Technology, Kharagpur, for their kind interest and encouragement.

Drs. R. K. Som and K. R. Parthasarathi of the Indian Statistical Institute gave useful suggestions to the author. Shri J. Saha, Librarian, ISI, drew his attention to important references, in the initial stages. Shri J. Roy of the Linguistic Research Unit of the ISI kindly placed one word-count at the disposal of the author.

A large number of colleagues shared the burden of the heavy statistical computations. Only the foremost among them can be mentioned by name : Sarvashree Ramdulal Chatterjee, Pareshnath Bhattacharya (PNB), (late) Amal Kumar Sengupta and Gobardhan Paul. Shri Rabindranath Mukherjee (RM) supervised many of these computations and also helped by making fruitful suggestions. This thesis could never have been completed without his deep interest and co-operation.

In addition to PNB and RM named above, Sarvashree Durgaprasad Bhattacharya and Suhas Chattopadhyay participated as subjects in the hand-writing experiment reported in Appendix 6.

Shri Khagendra Chandra Neogi did all the typing work; Shri Somes Saha prepared the mimeographed graphs and Shri Sanat Maity, those drawn in ink; Shri Manoharan De helped in comparing work; Shri B. K. Sinha prepared the photostat copies of the specimens of handwriting; Shri Manik Mitra operated the gestetner machine; and Shri Madhab Chakravarti was entrusted with the binding work. The author is glad to acknowledge the excellent co-operation received from all these persons.

He, of course, is responsible for all the deficiencies.

Indian Statistical Institute
Calcutta-35.
October 1965.

N. Bhattacharya

Contents

Chapter		Page
Chapter 0.	Abstract of Chapters 1-11 and Appendices 1-6	1 - 40
Chapter 1.	A Survey of Earlier Researches	41 -102
1.1	Introductory	41
1.2	Statistical Studies on Style	45
1.3	Statistical Studies on Languages	62
1.4	Information-theoretic Studies on Language	83
1.5	Indian Work in the Field	97
Chapter 2.	The Samples of Words	103 -141
2.1	Introductory Remarks on Coverage	103
2.2	The 'Word' and the 'Syllable'	105
2.3	Probability Sampling	108
2.4	Properties of Probability Sample Estimates	112
2.5	Interpenetrating Subsamples (IPNS)	117
2.6	Systematic Samples	123
2.7	Validity of Systematic Samples	126
2.8	Complete Count Figures for Poems	140
Chapter 3.	Word-length in Bengali — a Broad Survey	142 -186
3.1	The Material	142
3.2	Historical Trends in Word-length	167
3.3	Word-length in Different Types of Literary Works	171
3.4	Within Author Differences	176
3.5	Word-length Distributions — Other Aspects	184

Chapter		page
Chapter 4.	Words Within and Outside Conversations	187 -223
4.1	Introduction	187
4.2	The Material	191
4.3	Percentage of Conversational Matter (p_c)	210
4.4	Observations on $\bar{x}_c$ and $\bar{x}_o$	213
Chapter 5.	Randomness of the Word-length Series	224 -274
5.1	Introductory	224
5.2	Some Evidence of Deviations from Perfect Randomness	226
5.3	The Estimation of Autocorrelation Coefficients	229
5.4	The Autocorrelation Coefficients for Bengali Prose	242
5.5	Evidence from Short Bengali Passages	255
5.6	Some Investigation on English	263
Chapter 6.	The Form of the Word-length Distribution	275 -316
6.1	Introduction	275
6.2	Goodness of Fit Criteria	278
6.3	Inadequacy of Poisson Law for Bengali Prose	293
6.4	Poisson Law for Other Languages	298
6.5	Fitting Lognormal Distributions	302
6.6	The Goodness of the Lognormal Fit	311
6.7	Concluding Observations	315
Chapter 7.	A Classification of Bengali Syllables	317 -333
7.1	Introductory	317
7.2	Relative Frequencies of Different Types of Syllables in Bengali Prose	319
7.3	Relative Frequencies of Syllable-types in Bengali Poetry	323
7.4	Composition of Words in terms of Different Types of Syllables	327

Chapter		page
Chapter 8.	Sentence-length in Bengali Prose - (1)	334 -392
8.1	Introduction	334
8.2	Definition of Sentence-length	337
8.3	Probability Sampling of Sentences	338
8.4	Sentence-sampling	348
8.5	Sentence-length in Bengali Prose — a Broad Survey	350
8.6	Within Author Differences	370
8.7	Lognormality of Sentence-length Distributions in Bengali	375
Chapter 9.	Sentence-length in Bengali Prose — (2)	393 -427
9.1	Introduction	393
9.2	Tests of Randomness	395
9.3	Results of Tests of Randomness	408
9.4	Autocorrelations for Complete Works	414
9.5	Systematic Samples versus Probability Samples	415
9.6	Some Evidence on Homogeneity of Works	426
Chapter 10.	Correlation between Word-length and Word-frequency in Written English	428 -465
10.1	Introduction	428
10.2	Special Significance of the Correlation Coefficient	431
10.3	Empirical Values of the Correlation Coefficient	435
10.4	The Correlation Ratios	440
10.5	Limiting Values and Related Questions	444
10.6	Some Observations on the Joint Distribution of x and y	458
10.7	Concluding Observations	464

Chapter

		page
Chapter 11.	Intervals between Occurrences of the Same Word	466 -503
11.1	Introduction	466
11.2	Re-examination of Zipf's Material	469
11.3	Four Grammar Words of English	482
11.4	Some Grammar Forms in Bengali	492
11.5	Randomness of Series of Intervals	496

Appendix

Appendix 1.	Word-length in letters in Bengali Prose	504
Appendix 2.	Concentration Curve of Word-length and Sentence-length	511
Appendix 3.	On Word-length in Jane Austen's "Pride and Prejudice"	520
Appendix 4.	Relative Frequencies of Letters in Bengali Prose	528
Appendix 5.	The Rank-frequency Relation for Words in Bengali Prose	536
Appendix 6.	An Experimental Comparison between Speed of Writing in English and Bengali	544
References		571

Chapter 0 : Abstract of Chapters 1 -11 and Appendices 1 - 6

Chapter 1 : A survey of earlier researches .

0.1.1. Statistical studies on languages and literary styles have been carried out by many researchers during the past 100 years or so, and the subject is now emerging as a distinct branch of applied statistics. But although eminent linguists and students of literature have contributed to this field, many linguists and students of literature are sceptical about the value of the statistical approach.

0.1.2. This chapter attempts a broad survey of previous researches classified under three heads (i) studies on style, (ii) studies on statistical properties of languages and (iii) information-theoretic analysis. One section is devoted to Indian work in various lines.

0.1.3. Style is an essentially statistical concept, being based on whole works rather than on individual words or expressions, and it is desirable to have objective statistical style-indices. Most indices use distributions of word-length or sentence-length or the size, composition and diversity of vocabulary or metrical data for poetry. Among composite measures, one may note the readability indices like that due to Flesch (1948).

0.1.4. Applications of style measures are reviewed, including those to the chronology of Plato's or Shakespeare's works or to problems of authorship of Shakespeare's plays and the "De Imitatione Christi". The scope and limitations of such measures are discussed and also the statistical problems and methodology of using them, for instance, the problem of estimating size and vocabulary of an author from word

counts. It is emphasised that style measures may vary with the type of writing and with the age of the author. How far they are indicators of individual style has not been tested on any large scale.

0.1.5. Under the second head are discussed, among other things, the studies on relative frequencies of verbal units (phonemes, letters, syllables and words) and of pairs or triplets of letters and phonemes[1]; the Zipf law of word-frequencies and the probability models suggested for explaining it; the various economic principles underlying linguistic phenomena like the negative correlation between word-length and word-frequency; Ross' (1950) application of probability models to some problems of linguistic interest; and the applicability of probability models to texts or speech samples considered as sequences of words.

0.1.6. Probability sampling has scarcely been used for studies on languages or literary style. Subjectively chosen samples have often been used, and sometimes systematic samples of the non-probabilistic type. Yet statistical methods valid for strictly random samples were used in a few cases, without due reserve. A major objective of the present study is to use probability sampling and to justify non-probabilistic systematic samples as approximations to probability samples. The question of sampling error has been constantly kept in view. Standard errors were frequently not calculated in the detailed way : The technique of independent and interpenetrating networks of subsamples (IPNS) introduced by Mahalanobis (1946) was extensively used

1/ The relative frequencies of combinations of letters/phonemes are fairly stable within a language — this forms the basis for information-theoretic analyses of languages.

in the present investigation.

0.1.7. Information-theoretic analyses of actual languages was started by Shannon (1948, 1951). (The relevant theory of discrete communication on a noiseless channel has been summarised.) Shannon's work is described, in particular, the guessing experiment devised by him for estimating the entropy of a text. From the success with which subjects guessed the kth letter when k -1 preceding letters are given (k = 1,2,...,15,101), Shannon estimated that in ordinary literary English the entropy is about 1 bit per letter so that redundancy is nearly 75%. The estimate was confirmed by the more elaborate experiments of Burton and Licklider (1953). This means, in plain words, that English messages could be encoded with about one-fourth as many characters if all combinations of letters (and space) were possible and equiprobable.

0.1.8. The review also covers, among other things, the objective procedures of estimating entropy originally proposed by Newman and Gerstman (1952); the study by Miller and Friedman (1957) on the possibilities of abbreviating English texts without modifying rules of orthography; the inappropriate uses of entropy measures, especially by Herdan (1955a, 1956); and finally certain studies on relative economies of different languages i.e. of relative lengths of the same message expressed in different languages.

0.1.9. The final section on Indian work reviews Chatterjee's (1926) studies on Bengali on (i) historical trends in relative frequencies of 'tatsama' words[1/], (ii) relative frequencies of Persian words and

---

1/ i.e. Sanskrit words in relatively unmodified form.

(iii) relative frequencies of phonemes; Subba Rao's (1960) work on sentence-length in Kannada prose; word-counts on Bengali for pedagogic and other uses by Deb Chowdhury (1931) and by J. Roy and others of the Indian Statistical Institute; and Siromoney's (1963) study on the entropy of Tamil prose.

Chapter 2 : The samples of words

0.2.1. This chapter describes the material collected for word-length and related studies on Bengali, the methods of probability and systematic sampling adopted for such purposes, and the properties of estimates obtained from such samples.

0.2.2. Twentyeight works in Bengali prose were covered. These included seven novels of Bamkimchandra and eight by Tagore, and represent 100 years of modern Bengali prose, beginning with Vidyasagar's "Shakuntala" (1854). In addition,we studied three short essays, three short stories, twentytwo representative poems by Tagore and an extract from Michael M. Dutta's epic, "Meghanadabadha Kavya". Probability and/or systematic samples of words were drawn from each of the prose works; the poems/poetical pieces were generally subjected to complete counts.

0.2.3. Words were taken as printed, delimited by spaces. No attempt was made to count compounds of two words, say, as two words instead of one. The syllable-counts were based on the standard pronunciation of literary Bengali, which means the modes prevailing in learned circles in and around Calcutta. Some diphthongs, e.g., 'ai', 'oi', were regarded as similar to single vowel sounds, while others, like 'ea' and 'ia'

were treated as indicating two distinct syllables. Triphthongs etc., were split into different syllables on the basis of these rules.

0.2.4. Probability sampling was used for 24 of the 28 prose works. For each work, the desired number of lines (100 to 250) were selected strictly at random and with replacement, and words falling on sample lines formed the probability sample. Hyphenated words at the beginning (end) of sample lines were wholly excluded (included)[1]. Such use of line-clusters is quite efficient since intra-line correlations between lengths of words are small [vide Chapter 5]; unrestricted random samples (of words) would be relatively troublesome. The optimum design seems to be, on retrospect, the use of clusters of several consecutive lines.

0.2.5. Probability sampling has seldom been used by earlier workers, but may be profitably used where complete counts are unnecessary, and where at present subjectively selected passages are taken. Such subjective samples are also inferior to the 'systematic' samples used here [see below] since, although completely non-probabilistic, these 'systematic' samples are objective and behave like probability samples, to a close approximation. Actually, the two types of samples from the same work were pooled for purposes of statistical inference.

0.2.6. If n_i denotes the number of words on the ith sample line ($i = 1, 2, \ldots, k$), $n_i^{(r)}$ the number of r-syllabled words among these ($r=1, 2, \ldots$) and x_{ij} the length in syllables of the jth word on

---

1/ These sample lines were again used for drawing probability samples of sentences [vide Chapter 8].

on this i-th line ($j = 1,2,\ldots, n_i$), then we are mainly interested in the following ratio estimates :

$$p_r = \sum_i n_i^{(r)} \Big/ \sum_i n_i \quad \text{and} \quad \bar{x} = \sum_i \sum_j x_{ij} \Big/ \sum_i n_i$$

By referring to the theory of ratio estimates, it is shown that, excepting for the small proportions like p_7 **or** p_8, these estimates based on the probability samples have the large sample properties of consistency, asymptotic normality etc. In particular, the estimates $\bar{x}$ are approximately normally distributed with negligible bias, and the sampling variance is approximately given by

$$V(\bar{x}) = \frac{1}{k(k-1)\bar{n}^2} \sum_{i=1}^{k} \left(\sum_j x_{ij} - n_i \bar{x} \right)^2$$

where $\bar{n}$ is the sample average of n_i's.

0.2.7. The sample of k lines from any work was split into (usually) four independent and interpenetrating subsamples (SS) : SS 1 comprised lines numbered 1, 2, ..., $\frac{k}{4}$ in the order of selection; SS 2, those numbered $\frac{k}{4} + 1$, ..., $\frac{k}{2}$; and so on. Estimates based on subsamples were used for judging the extent of sampling errors. (The subsample estimates $\bar{x}$ generally possess the large sample properties of ratio estimates.)

0.2.8. The combined (i.e., all subsample) estimate $\bar{x}$ is close to the simple average of the four subsample estimates for all the **24** works. Since bias of ratio estimates is of order $\frac{1}{k}$, this confirms that even

the subsample estimates are practically unbiased. Similar agreement was found for the systematic samples and also for the estimates p_r from both types of samples, excepting for high values of r.

0.2.9. The systematic samples from the prose works were drawn by selecting, say, the 4th line from top of every odd-numbered page. (More than one such rule was frequently used for the same work.) Strict equalisation of the intervals between successive sample lines would have been time-consuming since the sampling fractions were small. For the short essays and stories, the sampling fractions were high, and every 3rd line (say) was selected in these cases.

0.2.10. The lines constituting the systematic sample from any work were split up into 4 interpenetrating subsamples (SS), SS 1 comprising the lines occupying 1st, 5th ... positions in the natural reading order, SS 2 those in 2nd, 6th ... positions, and so on. No use was made of any kind of random start, so strictly speaking, one cannot speak of sampling errors. Even if a random start were made, estimation of sampling errors would have been extremely difficult (theoretically impossible). Taking a "practical" view of the situation, sampling errors were assessed by the divergence between subsample estimates.

0.2.11. For 14 of the 28 prose works, both methods of sampling were used. The two sets of estimates showed satisfactory agreement in most cases. A few appreciable differences are rather to be expected when a large number of comparisons are made.

0.2.12. Four series of χ^2-tests[1] demonstrated the following:.

(i) the estimates (p_r and $\bar{x}$) from systematic samples are not significantly different from those from probability samples;

(ii) the sampling errors of the two sets of estimates are very nearly equal, apart from differences in sample sizes, and slightly larger than those for strictly random samples of the same size.

0.2.13. In the first series of χ^2-tests, χ^2 is applied, separately for different works to test the homogeneity of the word-length distributions from the subsamples of the probability sample; the second series does the same thing for the systematic samples from different works. The χ^2's in both series seem to have an upward bias. The third series compares the word-length distributions from the combined probability and the combined systematic samples from each of 14 works. The χ^2's of this series seem to have a downward bias. In the fourth series of tests, we compared the variability of the four subsample $\bar{x}$'s from the systematic sample with the standard error of combined $\bar{x}$ from the probability sample. The χ^2's showed a reasonable distribution of values. These findings have been discussed in detail. The broad conclusions are given in the foregoing paragraph.

0.2.14. The question of sampling errors cannot be completely answered even by the complete counts. Each poem/poetical piece was therefore

---

1/ and also fractile graphical analysis.

divided into two "parts", suitably near the middle[1/], and the divergence between "part" figures used to indicate the reliability of the complete count figures.

Chapter 3: Word-length in Bengali — a broad survey

0.3.1. This chapter examines the word-length distributions observed for different prose works and poems in Bengali. This reveals the historical trends in word-length and the appreciable within author variation in many cases; dimensional ideas are also obtained about word-length in different fields of literature.

0.3.2. Word-length has been measured in syllables. Word-length distributions are presented in detail, along with averages, s.d.'s and entropies. For the prose works, the distributions and averages are given by type of sample (probability/systematic) and also by sub-samples. The standard errors of averages were calculated for the probability samples; the s.e. for the pooled (probability plus systematic) sample average was assumed to be $\sqrt{\frac{n}{n+n'}}$ times the s.e. for the probability sample, where n, n' are the numbers of words in the probability and the systematic samples respectively. For the poems, the averages are given by "parts".

0.3.3. The averages ($\bar{x}$) for prose works depict a clear declining trend. Vidyasagar's "Shakuntala" (1854) and "Sitar Vanavas" (1860) show averages near 2.7. Both are written in the chaste sanskritised style abounding with compounds. Averages around 2.6 are seen in

1/ this was rather arbitrary in some cases.

earliest novels of Bamkimchandra, "Durgeshnandini" (1865) and "Kapal-kundala" (1866). But Bamkimchandra gradually moved towards a de facto colloquial style, and the averages for "Krishnakanter Will" (1878) and "Devi Chaudhurani" (1884) are as low as 2.35 (± 0.023) and 2.26 (±0.023)[1/]. Tagore employed the chaste style upto "Chokher Bali" (1903) where $\bar{x}$=2.37, and in "Chaturanga" (1916) where $\bar{x}$ = 2.32 (± 0.020); "Gora" (1910) used the colloquial style in conversations but the chaste style elsewhere and showed an average of 2.34 (± 0.021). Tagore's later works in the completely colloquial style show even lower averages — "Ghare Baire" (1916) : 2.09 (± 0.021), "Yogayog" (1929) : 2.17 and "Sheser Kavita" (1929) : 2.20 (± 0.023). Post-Tagore writers hardly show any further trend.

0.3.4. Averages for the 22 poems of Tagore show little evidence of time-trend. This is essentially because there is not much insistence that poetry should employ everyday language.

0.3.5. Among the poems, $\bar{x}$ varies from 3.35 for the lyric "Varsamangal" to 1.95 for "Virpurus", a poem for children, depending on theme, mood and meter[2/]. Poems with high $\bar{x}$ are usually on elevated topics, but the converse is not true. Chaste verbs and pronoun forms may be associated with low values of $\bar{x}$, and colloquial forms with high $\bar{x}$.

0.3.6. The range of $\bar{x}$ is narrower in prose fiction, about 2.1 to 2.5, excluding works in the now outmoded Sanskritised style of Vidyasagar and Bamkimchandra (early phase). Since conversational words tend to be

---

1/ The figures within brackets are the standard errors of the averages.

2/ The value of $\bar{x}$ cannot be appreciably below 2 in any non-trivial writing in Bengali.

shorter than other words, the averages tend to be lower for works with more of conversations, and especially for works like "Ghare Baire" written as speeches or thoughts of the leading character(s) so that even non-conversational matter is akin to conversation. The effect of the colloquial style has already become evident. But forms of verbs and pronouns alone do not matter much.

0.3.7. For essays, where conversations cannot appear, $\bar{x}$ ranges from 2.3 to 2.7. Historical novels and journalistic literature also tend to show somewhat higher values of $\bar{x}$.

0.3.8. Appreciable and significant within author differences are found among novels by Bankimchandra, Tagore and others. Word-length distributions cannot definitely indicate individual style in Bengali prose or poetry. The same has been found for sentence-length in Bengali prose [Vide Chapter 8].

0.3.9. Some earlier investigations seem to have created a false sense of optimism about the value of statistical indices as indicator of individual style. It is sometimes recognised that the style measures may vary with the broad field of literature, with the subject matter within a given field and also with the age of the author. But the situation seems to be even more complicated for Bengali prose, for some of the within author differences cannot be explained by such considerations.

Chapter 4 : Words within and outside conversations

0.4.1. Words in fiction were classified into 'conversational' and 'others' and this led to a better understanding of the between works

variation in word-length. 'Conversational' matter was defined as including only words actually uttered in conversations with persons present; soliloquies etc.. were excluded.

0.4.2. The proportion (p_c) of words coming under conversation is itself an interesting indicator of style. Also, since the writer has to make two distinct choices regarding the level of language to be employed, one for the conversation and another for the remaining narrative, it is desirable to study the lengths of the two classes of words separately.

0.4.3. Thus, p_c is about 26% for "Pather Panchali" and "Aparajita", but about 55% for "Devayan". The average word-lengths $\bar{x}$ are nearly 2.27, 2.27 and 2.14 respectively. The average lengths ($\bar{x}_c$) of conversational words are respectively 1.91, 2.00 and 2.09, and the averages ($\bar{x}_o$) for other words 2.40, 2.37 and 2.19. The variation of $\bar{x}$-values is in a sense due to the larger p_c in "Devayan"; for if p_c were the same as in "Devayan", while $\bar{x}_c$ and $\bar{x}_o$ were as observed, the $\bar{x}$-values would have been 2.13 for "Pather Panchali" and 2.17 for "Aparajita".

0.4.4. Word-length distributions and averages are presented for different works separately for the two classes of words and by type of sample. The estimates of p_c, $\bar{x}_c$ and $\bar{x}_o$ are given by subsamples. The estimates p_c have wide margins of error and do not possess the large sample properties of ratio estimates; but satisfactory estimates require very large samples, and even the rough estimates suffice for important conclusions.

0.4.5. The estimates of p_c vary continuously from about 5% in "Kshudhita Pasan" to 65 or 70% in "Laboratory". Among typical novels, "Visavriksha" shows p_c = 17% and "Sheser Kavita" 51%. Appreciable within-author differences are noticed in several cases, e.g., between "Visavriksha" and "Krishnakanter Will" (p_c = 31%). As shown in paragraph 0.4.3, these partly explain the corresponding variation in the values of $\bar{x}$.

0.4.6. The average $\bar{x}_o$ is significantly larger than $\bar{x}_c$ for most works, and evidently, on the whole.

0.4.7. A scatter diagram was plotted with each work represented by a point, the x-coordinate being the value of $\bar{x}_c$ and the y-coordinate that of $\bar{x}_o$. The broad picture is as follows : The works in chaste style — by Vidyasagar, Bankimchandra and Tagore (upto "Chokher Bali") — fall around the line $\bar{x}_o - \bar{x}_c = 0.35$, approximately, and indicate a declining time-trend in both $\bar{x}_o$ and $\bar{x}_c$. Works like "Gora" employing the colloquial style in conversations but the chaste style elsewhere seem to have progressed along the same declining trend. Only the $\bar{x}_o$-values fell further when the colloquial style began to be used throughout, and works like "Sheser Kavita" fall around a line with $\bar{x}_o - \bar{x}_c = 0.1$ or 0.15. Most of these works (e.g., "Ghare Baire") are written as thoughts or speeches of the leading character(s), so even the non-conversational matter is akin to conversation. For some of these, the difference $\bar{x}_o - \bar{x}_c$ is not significant.

0.4.8. The contrast between the two classes of words is further apparent from Chapter 5. For some work on English, see Appendix 3.

Chapter 5: Randomness of the word-length series

0.5.1. An analogue of a time-series is obtained if the words of a given text are replaced by their respective lengths and the lengths are read in the normal reading order. The present chapter studies the randomness of such word-length series in Bengali prose. Some work is also done on English. Word-length is measured in syllables for Bengali and in letters for English. Fucks (1954) has showed that for some works in German and English, the lengths (in syllables) of two consecutive words are approximately independent; the autocorrelation coefficient r_1 ranged from — 0.065 to + 0.013.

0.5.2. The standard errors of the average word-lengths ($\bar{x}$) based on probability samples of words from Bengali prose works [Vide Chapter 2] tend to be larger by 10% than the standard errors for strictly random samples of the same number of words. This points to the presence of positive autocorrelations in the word-length series. Since conversational words are shorter, on the average, than other words [Vide Chapter 4], fiction shows alternate runs of shortish conversational words and longish "other" words. This non-random feature is more pronounced when $\bar{x}_o - \bar{x}_c$ is larger[1]. That systematic samples behave like probability samples [Vide Chapter 2] points, on the other hand, to a kind of randomness of the word-length series.

0.5.3. For 16 works in Bengali prose, the probability samples of words were used for estimation of r_1; for four of these works, r_2, r_3, r_4 and r_7 were also estimated. In each case, four subsamplewise estimates

1/ This may lead to positive autocorrelations even if autocorrelations are zero within conversational matter or other matter.

were used for drawing inferences. Denote by n_i the number of words on the ith sample line. Then for estimating r_s (s = 1,2,3,4,7) we used the first s words of the following line to get n_i word-pairs. The joint distributions point to the approximate statistical independence of lengths of neighbouring words.

0.5.4. For the sample sizes used here the estimated autocorrelation coefficients may not be normally distributed to a close approximation. The estimates r_1 have a small negative bias. This is seen by comparing the simple averages of subsample estimates with the combined estimates. For tests of significance we applied the one-sided sign test on the subsamplewise r's. The negative bias renders this safe for practical purposes.

0.5.5. The coefficient r_1 is significantly positive for only six works, but it is evidently positive, on the whole; thus, all works give a positive estimate from the combined sample. A nearly unbiased estimate of the average r_1 for 16 works is obtained as 0.06. The coefficients r_2, r_3, r_4 and r_7 are also near zero. Perhaps the correlogram does not fall rapidly to zero; thus, the average of r_7 for 4 works is 0.06 and significant.

0.5.6. Seven passages were examined, two from "Visavriksha" and five from "Dristipat". Each had 200 to 250 words. These were found in a preliminary search for passages with high, medium and low averages of word-length. The overall average of word-length is about 2.47 syllables for "Visavriksha" and 2.37 for "Dristipat". The two passages from the former work showed the unusual averages of 1.99 and 3.08 and the five from the

latter showed averages between 1.95 and 2.72. This points to the presence of more or less conspicuous patches, in the texts, with high, medium or low averages of word-length.

0.5.7. The values of r_1 within these passages are, on the average, 0.06 for "Visavriksha" and 0.04 for "Dristipat". The significance of these is not quite clear. It is interesting to note that the values for entire works estimated from probability samples are somewhat higher, viz., 0.14 for "Visavriksha" and 0.07 for "Dristipat". The estimates of r_{10} are a little below zero but nonsignificant for the two "Visavriksha" passages.

0.5.8. A probability sample of words from "Pride and Prejudice" gave $r_1 = -0.15$[1/], and a systematic sample from "A Tale of Two Cities" gave $r_1 = -0.11$. Subsample estimates showed that both were significantly negative. Fucks (1954) had considered only one non-typical English work, "Othello" by Shakespeare, and found $r_1 = -0.02992$. Examination of some passages from "Pride and Prejudice" and "Othello" indicated that the distinction between letters and syllables is not very consequential.

0.5.9. The value of r_1 was practically the same for conversational word-pairs, non-conversational word-pairs and all word-pairs in "Pride and Prejudice". So the distinction between conversational and other words does not have any clear effect. The negative r_1 may be largely due to the tendency of shortish grammar words and longish content words to occur alternately. This tendency is much less clear in Bengali prose, where r_1 is about + 0.06; the contrast between conversational and other words may have contributed to make this positive.

1/ Vide Appendix 3 for an account of the material including the distinction between conversational and other words.

Chapter 6 : The form of the word-length distribution

0.6.1. This chapter examines the functional form of the distributions of words by length (x) in syllables (x= 1, 2, 3 ....) estimated for 28 works in Bengali prose[1]. For each work, the examination was based on the combined (all subsample) distribution; when both types of samples were available, the pooled (probability plus systematic) distribution was used. The small deviations from unrestricted random sampling [Vide Chapter 5] were ignored, but the conclusions were drawn with due reserve.

0.6.2. Elderton (1949) suggested the geometric distribution for certain data on English. Fucks (1955) stated that in eight of the nine languages examined by him[2], x-1 was approximately obeying the Poisson law[3]. The lognormal distribution was tried on some material on English by Williams (1956) and Herdan (1958); but here x was the number of letters. No goodness of fit measure was employed in these investigations.

0.6.3. The geometric distribution failed completely for the Bengali data. The Poisson model also gave a generally poor fit. Broadly speaking, the index $D = \sum_x |p_{obs} - p_{exp}|$ (where p denotes proportion of words of specified length) was about 0.15 or 0.20 for the older works in chaste style, about 0.30 for those with colloquial style in conversations and 0.35 or 0.40 for works in wholly colloquial style. The positive differences $(\bar{x} - 1) - s_x^2$ (variance) also showed the inadequacy of the model.

1/ Vide Appendix 1 for a study on word-length in letters.

2/ The exception was Arabic.

3/ Fucks (1955) examined one average distribution for each language, which is an ill defined concept. It is better to examine the distributions for individual works.

0.6.4. A re-examination of Fucks' data showed that the situation is not fully satisfactory for even the eight languages. For D=0.03 for Esperanto and 0.08 for German, but about 0.10 or 0.15 for the remaining six languages (and 0.31 for Arabic); and $(\bar{x} - 1) - s_x^2$ is clearly positive for 3 languages including Arabic, about zero for Esperanto and German, and clearly negative for the other languages. For the distributions for some individual works in English, German and Russian found in the literature, $(\bar{x} - 1) - s_x^2$ is usually negative. A few works show D=0.02 or 0.03; for the others D is 0.10 or 0.15 or even 0.30.

0.6.5. The lognormal curve was fitted to the discrete word-length distributions in two ways. In the first approach, denoted by LN(a), we supposed that the observed values of x, viz., 1, 2, ... represent the intervals 0 - 1, 1 - 2, ... of the underlying lognormal variate[1/]; in the second approach, referred to as LN(b), the corresponding intervals are 0-1.5, 1.5 - 2.5 ..... (It makes little difference if the intervals are 0 - $\sqrt{1\times2}$, $\sqrt{1\times2}$ — $\sqrt{2\times3}$, ...) The fitting was done by a modification of the method of quantiles.

0.6.6. The Poisson fit is best for "Sitar Vanavas", and not much worse than the lognormal fits for older works in the chaste style. But the lognormal fits are increasingly better for the later works. Average D for the 28 works is 0.293 for the Poisson fit, 0.137 for LN(a) and 0.110 for LN(b). For LN(a), D is about 0.15 or 0.20 for the older works but only 0.05 or 0.1 for the later ones; for LN(b), D fluctuates around

---

1/ This is the approach of previous workers.

0.1, roughly speaking, without any time-trend. So LN(a) gives a better fit for works in entirely colloquial style, while LN(b) seems to be better for the other works.

0.6.7. The Kolmogorov statistic (K) was used for judging the goodness of the lognormal fits, although, for several reasons, it gave an insensitive and conservative test in the present case. For LN(a), the K was about 5% for the earlier works but about 2% for the later ones; for LN(b), the values fluctuate around 3% without any time-trend [c.f. foregoing para 7].

0.6.8. The lognormal fits are thus fairly close, but samples being relatively large, goodness of fit tests give clearly significant results. Even K is significant (at the 5% level) for 14 works for LN(a) and for 12 works for LN(b). The χ^2-test used for LN(b) gave significant results for most works and evidently, on the whole.

0.6.9. The lognormal fit might be closer for conversational words or other words than for all words.

Chapter 7: A Classification of Bengali Syllables

0.7.1. This chapter introduces a simple classification of Bengali syllables and studies the relative frequencies of different classes in Bengali prose and poetry.

0.7.2. Open (vowel-ending) syllables without diphthongs were called 'type A' syllables. The remaining 'type B' syllables are sometimes split into two : closed (consonant-ending) syllables, called type B_1

and open syllables with diphthongs, called type B_2. This latter distinction is not important, but the two broad types are each relatively homogeneous and type B syllables are longer than type A syllables. So the total number of syllables cannot adequately measure the length of a word
Bengali /: one should state the numbers falling under the two types separately.

0.7.3. About one-third of the syllables in Bengali prose fall under type B as against nearly two-thirds in English prose. This partly bridges the gap between the average word-lengths, about 2.25 syllables in Bengali prose and about 1.45 in English prose. But type B syllables in English are not quite homogeneous in length, and some of them are longer than Bengali type B syllables. So such approaches can only give rough indications and it is simpler to measure word-length in letters or phonemes. But data on syllable-types are of interest by themselves.

0.7.4. Relative frequencies of the different types were estimated for 16 prose works, and all the short essays, short stories and poems/poetical pieces mentioned in Chapter 2. Subsamplewise estimates were prepared for prose and 'partwise' estimates for poetry.

0.7.5. The percentage of type A syllables was relatively stable, between 62% and 72% in Bengali prose, although many between works differences were significant. So the total number of syllables is nearly satisfactory for between works comparisons of word-length. The percentage of type B_2 syllables is about 5% in Bengali prose and 3% in poetry.

0.7.6. The percentage of type A syllables vary markedly, from 65 to 90, among the poems/poetry pieces and this variation seemed to be related to the meter employed. A lyric by Satyendranath Dutta, viz., "Sindhutandab", not studied here, employs a meter where the two types of syllables must alternate.

0.7.7. Two-way distributions are formed showing frequencies of words according the numbers of type A and type B syllables comprising them. Some data are also given on the percentages of different types of syllables in different positions within words of specified lengths.

0.7.8. The simple Markov chain could not fit the sample texts in Bengali prose considered as sequences of 3 or 4 types of elements, viz., the 2 or 3 types of syllables and the gap **g** between words.

Chapter 8: Sentence-length in Bengali prose -(1)

0.8.1. Chapters 8 and 9 report on some sentence-length studies almost wholly confined to Bengali prose. Sentence-length was measured in terms of the number of words. The studies are parallel to those on word-length covered in Chapters 2 to 6, but on a smaller scale. Chapter 8 discusses (i) the method of probability sampling used for sampling sentences, (ii) the sentence-length distributions thrown up for Bengali prose, which give dimensional ideas about sentence-length in different types of texts, (iii) the within author variation in sentence-length showing how far sentence-length can indicate an author's individual style, and (iv) the lognormality of the distributions for Bengali prose.

0.8.2. Previous studies on sentence-length are reviewed. Williams (1940) showed that sentence-length distributions are approximately lognormal, but he did not use any goodness of fit criterion; his investigation was also on a small scale. Yule (1938) and Subba Rao (1960) concluded that sentence-length is an indicator of individual style in prose without really showing that within author variation was negligible. Probability sampling has seldom been used and statistical methods valid for unrestricted random sampling have sometimes been used uncritically without due reserve. Apparently Yule (1938) could not find any convenient method of probability sampling.

0.8.3. Probability samples of 200 or more sentences were selected from each of 19 works in Bengali prose. A number of lines were selected at random, with equal probability and with replacement, from each work[1/]. The sample included sentences having termination marks, that is, ending on these sample lines; when a sample line gave at least one sentence ending on it, we also included in the sample, the sentence(s) ending on the immediately preceding line which had at least one sentence ending on it and also the sentence(s) ending on the immediately following line which had at least one sentence ending on it.

0.8.4. This gives equal probability of inclusion to almost all sentences in the work, but there is an element of clustering, and the clusters are overlapping. The sample of sentence-clusters from any work was split up into four independent and interpenetrating net-works of subsamples (IPNS), and the IPNS technique was adopted for assessing sampling errors.

1/ For most of the works, the lines selected for probability sampling of words (vide Chapter 2) were used.

0.8.5. Sampling properties of the estimates are discussed. The estimated averages of sentence-length are practically unbiased, and possess the large sample properties of normality etc., of ratio estimates, though not very clearly at the subsample level.

0.8.6. Systematic samples of sentences were taken from three of the 19 works subjected to probability sampling. These samples could be used as approximations to probability samples [vide Chapter 9]. In addition, three short essays in Bengali and certain extracts from Bengali and English novels were completely counted for studying sentence-length.

0.8.7. Author's punctuation was usually accepted. Semi-colons and colons were not regarded as termination marks, excepting where a colon introduced a speech with two or more sentences[1]. Words were taken as printed, demarcated by spaces.

0.8.8. The average sentence-length is 7.25 for "Kavi Shri-Ramakrishna", and about 9 for several others; 11 or 12 seems to be the modal range in Bengali fiction; most essays and novels with elevated discussions (e.g., "Gora") show higher averages, say, 15 or 16. Unlike the averages of word-length, the sentence-length averages did not show any time-trend. Some comparisons are made with previous results for other languages.

0.8.9. Since the sentence-length distributions were approximately lognormal, [see below], the logarithmic standard deviation of the fitted distribution was used for drawing inferences. This, in fact,

---

1/ Semicolons were too often used in Vidyasagar's works; for these works some rough estimates were also obtained after changing the inappropriate semi-colons into full-stops.

indicated the C. V. of the distribution, and together with the average, practically summarised the information contained in the sentence-length distribution. Works were classified jointly on the basis of these two characteristics. The two characteristics seemed to be relatively independent and the C. V. showed within author variation unexpectedly in certain cases where the average did not.

0.8.10. The average (also C.V.) of sentence-length showed appreciable, within author differences. Thus, the averages were 15.54 for "Gora", 12.43 for "Chaturanga," Parts 1 and 2, 11.02 for "Sheser Kavita", and 10.29 for "Yogayog". Again, "Pallisamaj" and "Pather Dabi" had very similar averages of both word-length and sentence-length, but the C.V. of sentence-length seems to be higher for the latter work. These differences might be attributed to changes in style with the age of the author, or with the subject-matter of the work, say; but the usefulness of sentence-length as style-indicator appears to be limited in Bengali prose [cf. Chapter 3].

0.8.11. The lognormal distributions were fitted separately for 20 works, by the method of quantiles, regarding the observed lengths 1, 2, 3, ... as intervals 0-1, 1-2, 2-3,.., of the underlying variate. The Kolmogorov statistic was used for testing the goodness of fit: That parameters were estimated from the sample rendered the test somewhat conservative[1]. But the K-test gave clearly non-significant results for all the 20 works and also, on the whole.

1/ The discreteness of the variate and the small deviations from strictly random sampling [vide Chap. 9] has small effects in the same direction.

0.8.12. The ogives on log-probit scale were convex to the horizontal axis for most of the 20 works. A partial explanation seemed to be the heterogeneity of the population of sentences. For example, "conversational" sentences are shorter, on an average, than "other" sentences. This might lead to an excess of very short and also very long sentences (when all sentences are considered together) compared with the log-normal hypothesis.

Chapter 9: Sentence-length in Bengali Prose - (2).

0.9.1. If the lengths $x_1, x_2, \ldots, x_n$ of different sentences of a prose work are recorded in the natural reading order, one gets what may be called a sentence-length series. The randomness of such series for some Bengali (and English) works has been examined in this chapter mainly by means of non-parametric tests [cf. Chapter 5].

0.9.2. Randomness of sentence-length series has not been studied by earlier workers. Yule (1938) stated that the series seemed to be autocorrelated, creating difficulties in sampling and in statistical inference. But Williams (1940) and Subba Rao (1960) ignored this possibility and implicitly assumed that the series are nearly random.

0.9.3. This chapter also shows the validity of the systematic samples of sentences, used in Chapter 8, as approximations to probability samples [cf. Chapter 2].

0.9.4. The sentence-length series examined were obtained from (i) first two parts of Tagore's four-part novel, "Chaturanga", (ii) one chapter from Tagore's "Sheser Kavita", (iii) three short

essays by Bankimchandra and Tagore and (iv) two extracts from Jane Austen's "Pride and Prejudice". Two-way distributions pointed to the approximate statistical independence of consecutive sentence-lengths.

0.9.5. In addition to fractile graphical analysis, four nonparametric tests were applied to examine the randomness of each series : (i) Mann's test of trend based on Kendall's rank correlation coefficient between t and x_t [1], (ii) the Wald-Wolfowitz test for circular autocorrelation coefficients r_1, r_2, ..., which like the two following, is not really sensitive to trend, (iii) Wallis-Moore test based on the total number of turning points and (iv) the Wallis-Moore χ^2-test based on the observed and expected distributions of phase-lengths. All these procedures test are described in detail and their properties discussed. Attempts are made to combine the tests for the different series.

0.9.6. These tests point to the approximate randomness of the sentence-length series examined. Significantly positive or negative coefficient values of the rank-correlation are noticed in a few cases. But, on the whole, the series do not show any rising or falling trend. The two Wallis-Moore tests give non-significant results for all the series and also on the whole, so the series do not show any significant oscillations. The average value of the circular autocorrelation coefficient r_1 is 0.07 when the eight series are considered together, but this is significantly positive. The estimates for individual series vary from

---

1/ This test was applied on the shorter series formed by the average lengths of sentences in mutually exclusive groups of 3, 5 or 10 consecutive sentences.

below 0 to about 0.2. The coefficients r_2, r_3, r_6 and r_{10} are also close to zero; but they do not suggest any clear shape of the correlogram. However, r_{10} is nowhere significant. Possibly the type of text has some influence on these values.

0.9.7. The probability samples of sentences [Vide Chapter 8] were not used for estimating such autocorrelation coefficients based on entire works : these samples were too small for the purpose, and might also give somewhat biased results. The sample from "Krishnakanter Will" gave $r_1 = 0.185$, but the four subsample estimates diverged too much to give any precise idea of the true value.

0.9.8. Systematic samples of sentences were drawn from three works in Bengali prose subjected to probability sampling also. The systematic samples were drawn by selecting, say, the 4th line from top of every alternate page, and noting the lengths of sentences ending on these selected lines. Four such rules were adopted for each work, such that when all sample lines were considered together, the first gave lines occupying 1st, 5th, 9th .. positions in the natural reading order, the second those in positions 2, 6, 10, ..., and so on. This gave four independent and interpenetrating subsamples of the sample of sentences. The divergence between subsamples was assumed to be indicating the sampling error of the systematic samples.

0.9.9. Fractile graphical analysis and the two-sample Kolmogorov test indicated close agreement between the sentence-length distributions from the two types of samples. The sampling errors also were roughly equal, except for differences in sample size.

0.9.10. Some evidence is finally presented on the homogeneity of a few works in respect of the distribution of sentence-length.

Chapter 10: Correlation between word-length and word-frequency.

0.10.1. Word-counts on sufficiently long texts or speech-samples generally reveal a negative correlation between the lengths (x) of words in terms of syllables or phonemes or letters and their frequencies of occurrence (y) found in the count. The phenomenon is ascribed to the forces of linguistic evolution which set up some causal interdependence between x and y.

0.10.2. Zipf (1949) discussed the economic value of this negative correlation, demonstrated it for some material on English and Latin and stated that this correlation has been observed in a variety of languages. Herdan (1956) presents some material for English, German and Russian. Herdan (1958) fitted the regression $\bar{y}_x = ax^{-b}$ to some material on English with x as the number of letters; b was nearly 2.4. Miller, Newman and Friedman (1958) observed that the frequencies of content words[1] are "relatively independent of length", but the correlation is pronounced for function words[2]. All these studies were confined to the regressions of y on x or of x on y.

0.10.3. Correlation coefficients (r) and ratios (η) are presented in this chapter for the following word-counts on written English, measuring word-length in letters : (a) the Miller-Newman-Friedman (1958) count of 36299 words (5537 word-types), of which 14877 were content

---

1/ Nouns, verbs, adjectives plus most of the adverbs.
2/ Other than content words.

words (5180 word-types) and 21422 function words (357 word-types); (b) the Dewey (1923) count of 100,000 words, concretely, the list of 1027 particular words (78634 word-tokens) and that of 1132 root words (87380 word-tokens) occurring more than 10 times; and (c) Yule's (1944) counts on about 4000 nouns (about 1000 word-types) in each of four works by John Bunyan. The first two counts are based on representative works. The first used the particular word concept and the third, more or less, the root word concept.

0.10.4. Suppose on has n words (i.e., letter-combinations) devoid of meanings, having lengths $x_1, x_2, \ldots, x_n$, and n concepts or meanings to be used $y_1, y_2, \ldots, y_n$ times in a text. For random matching of words and meanings the text-length would, on the average, be $n\bar{x}\bar{y}$. In general, the text-length is $n\bar{x}\bar{y}\,(1 + rC_xC_y)$, where r is the correlation coefficient between x and y, and C_x, C_y the coefficients of variation.

0.10.5. Thus, although both regressions are appreciably curved, r is more meaningful than η, for it indicates the effect of the correlation on the length of the text. But for the observed marginal distributions of x and y, r cannot attain the limits ± 1, nor text-length the corresponding limits $n\bar{x}\bar{y}\,(1 \pm C_xC_y)$, for any matching of words and meanings. The actual limits were found by considering the most economic (uneconomic) matching where the jth most frequent word is matched with the jth shortest (longest) word, for $j = 1, 2, \ldots, n$.

0.10.6. This matching model is unrealistic, for it ignores the fact that related concepts are usually matched with structurally

related words, like 'go', 'goes', 'gone' etc. So the model is more appropriate for root words than for particular words. This distinction between root words and particular words would be really important for more inflected languages.

0.10.7. Values of r, η_{xy}, η_{yx} and $(1 + rC_xC_y)$ are given for the different word counts, and for the corresponding hypothetical most economic and most uneconomic matchings. These may be biased, being based on the section of relatively frequent words. The ultimate interest lies in values from infinitely large counts where even the rarest words are represented with their true relative frequencies. But these true values can hardly be defined precisely or meaningfully excepting perhaps for function words[1].

0.10.8. Some light was obtained by calculating r, η_{xy} etc., for the different counts, and for the associated hypothetical matchings, after excluding words upto different values of y. There was close agreement between estimates for particular words from corresponding sections[2] of the first two counts; the estimates for the Dewey list of root words were also similar. So the estimates from the complete Miller-Newman-Friedman count, upto y = 1, are the best available from this study for all root or particular words. The estimate of r is -0.14, and its attainable limits -0.20 and +0.28. The value of $(1 + rC_xC_y)$ is 0.62, and its limits are 0.46 and 1.76. The two correlation ratios are a little above 0.3. Since the estimates change systematically as rarer words are included, even these estimates may be appreciably biased.

1/ The results obtained for function words may not be far from the true values.

2/ The sections $y \geq 3, 4, \ldots$ in the Miller-Newman-Friedman count of 36299 words correspond, respectively, to the sections $y \geq 10, 12, \ldots$ in the Dewey count of 100,000 words.

0.10.9. Such results are also given for content words, function words and nouns. For content words and function words, the estimates do not change appreciably when rarer words are included. So the true values of r may not be far from -0.20 for content words and -0.30 for function words; the averages of the two correlation ratios are about 0.25 and 0.5 respectively. In all cases, the values for the hypothetical matchings are also discussed.

0.10.10. It is shown that the conditional distribution of y given x sometimes follows the Zipf law, approximately. Also, log log $\bar{y}_x$ seems to be linearly related to log x for content words or all words, though not for function words; but log $\bar{y}_x$ is nearly linearly related to log x for either content words or function words, but not for all words.

0.10.11. Finally, observations are made on the studies by Milller and Newman (1958) and Miller, Newman and Friedman (1958) on the length-frequency correlation and its relation to the Zipf law. It is also suggested that the length-frequency correlation be studied for other languages, with conversational material and using other measures of word-length.

Chapter 11: Intervals between successive occurrences of the same word

0.11.1. Suppose one numbers the different word-positions of a text serially in the natural reading order. If a given word occupies the positions $i_1, i_2, \ldots, i_f$, then $x_1 = i_2 - i_1$, $x_2 = i_3 - i_2$, ..., $x_{f-1} = i_f - i_{f-1}$, are the intervals between successive occurrences of the word[1/].

1/ Perhaps i_1 might be included as the first interval.

0.11.2. Zipf (1949) studied the length distributions of such intervals for rare words in James Joyce's "Ulysses", employing the Hanley (1937) word-index. Intervals were measured in pages. Each word gave only a few intervals, so Zipf pooled intervals of all words with the same frequency f. For 14 values of f between 5 and 24, Zipf found the approximate relation $n_x = a_f x^{-b_f}$, where n_x is the frequency of intervals of length x (x=1,2,3,...) and a_f, b_f are constants; also, b_f was close to 1 in most cases. Zipf also felt that the distributions of $x_1, x_2, \ldots, x_{f-1}$ were more or less the same.

0.11.3. Herdan (1956) found that for the very frequent Russian grammar form К (= 'to' etc.) the interval-distribution was very nearly exponential, to be strict, geometric, i.e., $n_x = ab^x$, since the variable is discrete. This suggests that the probability of the given word occurring is a constant (p) for all word-positions and the different positions are filled up independently and at random.

0.11.4. This chapter examines such interval-data for some English and Bengali words/word-groups by more rigorous statistical methods. Zipf's procedure for establishing his relation was clearly defective and his material demanded closer scrutiny. While Zipf's relation for rare is linear on double-log scale, Herdan's, for a single frequent Russian grammar form, English words is linear on semi-log scale. We examined the applicability of the geometric distribution to some frequent grammar words/word-groups in English and Bengali. Some work was done on the randomness of the interval-series $x_1, x_2, \ldots, x_{f-1}$. Such studies may throw light on the applicability of probability models to texts/speech-samples considered as sequences of words.

0.11.5. For f=5,15 and 24, we prepared the interval-distributions examined, but not presented, by Zipf, and found that the relation between log n_x and log x is appreciably curved if the whole range of x is considered. Zipf found a linear relation because he considered the range x =1 to 21 (50 in one case) only. The Zipf relation is also a priori unsuitable, for if b_f is close to 1, the distribution does not have a finite mean. Graphs suggested the geometric distribution, but even for this the fit was far from adequate — the Kolmogorov statistics were highly significant. Possibly as the context changes, the occurrence-probabilities of rare words change. Actually, short intervals were rather too frequent. For f=15 or 24, observed frequencies were rather too high upto x=10, roughly; for f=15, the geometric law seemed to be adequate for the distribution truncated at x=11. For f=5, however, the region of poor fit extended upto about x=35; the K-distance was 13%.

0.11.6. We counted the lengths in words of all intervals between successive occurrences of four frequent grammar words ('the', 'to', 'and' and 'of') in the English novel, "Pride and Prejudice" (Chaps.1-9). The geometric distribution seemed to be appropriate and the fit was better than for rare words in "Ulysses"; the K-statistic was only about 6 to 8%. But two of the four K's were significant.

0.11.7. We made similar interval-counts on (i) the word 'se (=he/she), (ii) the word 'sei' (=that) and (iii) the 'se' class of words considered as one word in the Bengali novel, "Gora", by Tagore. The Zipf relation again failed, and the geometric law was clearly better,

but even there the K-distances were about 6 to 9% and significant. Here again, short intervals were rather too frequent compared with the theoretical distribution.

0.11.8. The seven interval-series mentioned in paras 0.11.6-7 were subjected to fractile graphical analysis and four other non-parametric tests of randomness, viz., the Wallis-Moore test based on number of turning points, the Wallis-Moore χ^2-test based on the distribution of phase lengths and Mann's test of trend using Kendall's rank correlation coefficient between the serial number of the group of intervals and the mean/variance of the interval-group. The results were generally non-significant, demonstrating the approximate randomness of the series. This justifies the pooling together of all intervals of the same word for preparing interval-distributions.

0.11.9. We compared the separate frequency distributions of the four successive intervals x_1, x_2, x_3 and x_4 of the 898 words occurring 5 times in "Ulysses". The distribution of x_1 seemed to be significantly different from the others — this may be due to changes in the subject-matter within "Ulysses". We also examined the correlation coefficients between different pairs of intervals x_i and x_j for 97 words each occurring 15 times or 33 words each occurring 24 times in "Ulysses". The correlations are near zero and generally non-significant.

Appendix 1: Word-length in letters in Bengali prose

0.12.1. This appendix presents the distributions of words by length in letters estimated from the probability samples of words from "Visavriksha" and "Sheser Kavita" (vide Chap.2). Some conventions had to be adopted; thus, the double consonants like 'bb' were counted as two letters. The estimates for each work are given for each of the four subsamples comprising the probability sample as well as for the combined sample.

0.12.2. The average of word-length is around 5.84 letters in "Visavriksha" and 5.08 letters in "Sheser Kavita"; and the s.d.'s are about 2.57 and 2.27 respectively. The average in terms of letters is nearly 2.35 times the average in terms of syllables; for s.d, the ratio is about 2.4.

0.12.3. Ogives on log-probit scale indicated that the word-length distributions could be fitted fairly well by lognormal curves, especially by the LN(b) approach (vide Chap. 6).

0.12.4. The joint distributions of words by length in syllables and length in letters show close correlation between these measures of word-length. The correlation coefficient is near 0.9, and the regressions appear to be sensibly linear. The findings based on length in syllables should be more or less unaffected if word-length is measured in terms of letters; the same is probably true of word-length measured in phonemes also.

Appendix 2: Concentration curves of word-length

0.13.1. The definition of the classical concentration curve is extended to the case where the size-distribution is discrete, as that of word-length (Bhattacharya and Mahalanobis, 1964). If p_r is the proportion of cases where the variate x assumes the value x_r (r=1,2,...) — where x_1, x_2, ... are in increasing order — the curve is obtained by plotting $\sum_{j=1}^{r} p_j x_j / \bar{x}$ against $\sum_{j=1}^{r} p_j$ for r = 0,1,2,..., and joining the points successively by straight lines[1/]. The area of concentration thus obtained bears the usual relation to the Gini mean difference.

0.13.2. Halfsamplewise concentration curves of word-length in syllables are presented for selected works.

0.13.3. The common reader is sensitive to comparatively small differences in $\bar{x}$. First, the reader is sensitive to the presence of long words, and the proportions of polysyllabled words increase rapidly with $\bar{x}$. Second, when looking at random across a page, the reader tends to select words with probability proportional to length, so the length-distribution of selected words is given by $p'_x = x p_x / \sum x p_x \; (x = 0, 1, 2, \ldots)$ and p'_4, p'_5, etc. are larger than p_4, p_5 etc.

Appendix 3: On word-length in Jane Austen's "Pride and Prejudice"

0.14.1. A probability sample of 1772 words and a systematic sample of 1768 words were chosen, by methods described in Chap.2, from "Pride and Prejudice", Chaps.1-32, pp.1-199. Each sample comprised 4 interpenetrating subsamples. Word-length was measured in letters. This

1/ It is assumed that the average $\bar{x}$ exists.

appendix presents the word-length distributions and reports on some studies based on this material; the study on randomness of the series of word-lengths has been reported in Chap.5.

0.14.2. Homogeneity χ^2-test, applied in the same manner as in Chap.2, Section 2.7, (and also the Kolmogorov test) showed that the two methods of sampling gave very similar distributions of word-length with similar magnitudes of sampling error, and that both methods were more reliable than unrestricted random samples of the same size. (This letter is to be expected since the word-length series has a negative autocorrelation r_1, as shown in Chap. 5.)The agreement between the two methods of sampling was also seen from separate estimates for conversational and other words.

0.14.3. Nearly 45% words were conversational. Average lengths in letters were about 4.1, 4.67 and 4.41 for conversational words, other words and all words, respectively. The difference between conversational and other words is clearly significant.

0.14.4. The Kolmogorov-Smirnov test indicated the homogeneity of the word-length distributions estimated from the systematic sample from four "parts" of the work, viz. pp. 1-50, 51-100, 101-150 and 151-199.

0.14.5. The ogive on log-probit scale is curved even when the LN(b) approach (vide Chap. 6) is adopted. The lognormal fit should be better if conversational and other words are treated separately.

Appendix 4: Relative frequencies of letters in Bengali prose

0.15.1. The probability samples of words from "Visavriksha" and "Sheser Kavita" were used for getting subsamplewise and combined estimates of the relative frequencies of Bengali letters[1].

0.15.2. The vowels form about 46% of letter-occurrences, and the most frequent letters are vowels — 'অ' (about 15%), 'আ' (10-11%) and 'ি' (9-10%). However, over 2% of the letters are the silent* 'অ'. Percentage frequencies of other letters range from near 0 for 'ঋ', 'ঞ', 'ঙ', 'ঢ', 'ঌ', 'ঢ়' and 'ঐ' to about 7 for 'র' and 'ই'. Many interesting results are found. Thus, the aspirates tend to be less frequent than the corresponding unaspirates and the long vowels than the short ones.

0.15.3. There are appreciable differences between the relative frequencies for the two works, but these may be explained by the higher proportion of 'tatsama' words (Sanskrit words, more or less, unmodified) in "Visavriksha".

Appendix 5: The rank-frequency relation for words in Bengali prose

0.16.1. Two word counts on Bengali prose were studied here, (i) the count of all the 126359 words in "Gora" giving 15105 word-types made by J. Roy of the Indian Statistical Institute, and (ii) the count of 4456 words (1840 word-types) in the first story of "Chacha-Kahini" made by the author. Each inflected form was kept separate.

0.16.2. The rank-frequency relations were examined in the manner of Zipf (1949). Deviations from the Zipf law were noticed. Thus, the

---

1/ The material is the same as that employed in Appendix 1.

'standard curves' showed top-downward concavity for low values of the rank; also, the slope was nearly -0.8 for "Chacha-Kahini" but slightly steeper than -1 for "Gcra". But such deviations had been observed by Zipf and others and could be attributed to various factors.

Appendix 6: An experimental comparison between speed of writing in English and Bengali.

0.17.1. An experiment was carried out for comparing the time required for conveying the same message in written English and written Bengali. The experiment was on small scale and not strictly scientific, for unavoidable reasons; but it does indicate that the time requirements are nearly equal for the two languages.

0.17.2. The syllabic system of writing for Bengali has led to the impression that the current Bengali script is far less efficient than the roman script. This seems to be incorrect so far as handwriting is concerned. The romanised script proposed by Chatterjee (1935), has, however, other important advantages.

0.17.3. Five subjects each wrote 27 pairs of passages, each pair comprising an English passage and its Bengali translation or vice versa. All the passages were short, between 82 and 383 (English) words, and fatigue or practice showed negligible effects. A simple experimental 'design' was adopted. The subjects were asked to put in equal efforts when writing the two passages of each pair, keeping each letter legible. Two of them wrote as fast as they could, sacrificing quality, while the others wrote fairly good hands at fair speed.

0.17.4. Six of the original passages were from Tagore's Bengali novel, "Sheser Kavita", six from Nehru's "Autobiography" in English, five were English newspassages emanating from Reuter etc., and ten were English[1] passages set for translation into Bengali in a public examination. The translations were to all appearances faithful.

0.17.5. The Bengali : English ratios of time requirements varied to some extent between the five subjects. The average was about 99% for the passages from the "Autobiography", 105% for the news-passages and 104% for the examination passages. These represent journalistic and other factual matter, where translation is easy; and the ratios may not be significantly different from 100%. But "Sheser Kavita" uses highly idiomatic Bengali, which is translated with great difficulty, and the average ratio is well below 100%, viz. 84%. Probably the picture would be reversed if highly idiomatic English passages were chosen, for which translation into Bengali is equally difficult.

0.17.6. The small coverage of types of passages and the small sample of subjects are not major limitations of the experiment. True, all the five subjects were Bengalees; but all were graduates writing English more frequently than Bengali, and on an average, they wrote 32 English words per minute. Theoretically, the major limitations are :(1) The subjects had to equalise their efforts when writing the two passages of a pair in a completely subjective manner. (2) Nothing could be done to ensure that the two passages were written equally well and in letters of equal size[2].

0.17.7. It is suggested that similar studies be carried out for comparing time or energy requirements for conveying the same information in different languages (or scripts) by speaking, writing, typing, printing etc.

1/ Some of these were translations from Bengali originals.

2/ In actual practice, the equality of quality and size of letters seems to have been achieved, approximately, although such equality can hardly be defined.

Chapter 1: A Survey of Earlier Researches

1.1.1. Introductory: Linguistics has reached an advanced stage of development. It has studied in great detail the grammatical structure of different languages, and the evolution of languages with time (Jespersen, 1922 ; Taraporewala, 1951). The methods of linguistics are sometimes statistical, in a remote sense, if 'statistics' stands for systematic observation and assimilation of large volume of data. But so far as quantitative statistical methods are concerned, the science of linguistics has made little use of them.

1.1.2. The same may be said of literary criticism. (Of course, this has not attained the status of a science.) Even to-day, the scholarly critic seeks little help from the statistician when discussing the styles of different authors.

1.1.3. Statistical studies on languages and literary styles have been carried out from time to time by individual workers during the past 100 years or so. The frequency of such studies has increased in recent years under the influence of pioneers like G. K. Zipf, G. U. Yule and C. E. Shannon. A considerable body of literature is now interspersed in various books and journals. "Modern" statistical methods are being used with increasing frequency, and interesting statistical problems are continually being thrown up in these researches. The subject is, in short, emerging as a distinct branch of applied statistics.

1.1.4. These studies include many by eminent linguists or students of literature, who used statistical methods as aids in their

investigations. Thus, Forstemann studied the relative frequencies of phonemes in different languages and also used a statistical index of the divergence between the distributions of phonemes in two given languages (vide paras 1.3.1 — 1.3.3 , below). One may also mention the researches on the chronology of Plato's or Shakespeare's works (vide Section 1.2 below), and those by Chatterjee on the Bengali language (vide Section 1.5 below). Even Ross, a severe critic of many statistical studies, himself carried out several investigations (Ross, 1950) besides some information - theoretic analysis (Bell and Ross, 1956), this latter being of apparently little linguistic interest.

1.1.5. But there is no general agreement as yet about the value of these statistical researches, or in other words, about the role which statistics should play in its application to language (Herdan, 1956, Chap.1). The statistician is apt to regard language as just another object, suitable for quantitative investigations, no matter whether the results of such investigations are of sufficient interest to the linguist. Consider, for example, the finding that lengths of consecutive words are practically independent in German texts (Fucks, 1954). If the linguist considers these uninteresting or irrelevant, it is for him to widen his interests and acquire the statistical outlook. Some linguists, on the other hand, would like to apply statistics as an auxiliary tool in the very few linguistic problems where probability models seem to be applicable (Ross, 1956).

1.1.6. Admittedly, the statistician without sufficient knowledge of the field of application runs the risk of making a blind application of statistical techniques. This has actually been the case with several investigations strongly criticised by linguists. One may go through, for example, Gibson's criticism (Gibson, 1962, pp. 140- 7) of the statistical evidence in favour of the theory that Marlowe was Shakespeare. But as Yule says (Yule, 1944, p. 7) linguists without adequate knowledge of statistics have also made grave mistakes in trying to handle quantitative linguistic data. So, "neither the statistician without training in linguistics, nor the linguist without training in statistical methods is properly equipped for work in this field". In any case, such instances of misuse cannot detract from the scope and merit of the statistical approach.

1.1.7. One serious difficulty in bridging the gap between the linguist and the statistician, is the unacceptability to linguists of certain probability models put forward by Yule (1944, pp. 48-49) and others. These models conceive of texts or samples of speech as realizations of stochastic processes, words being chosen, one after another, following probability rules. Ross (1956) raises vehement objections to such views of language : he seems to regard speech or writing as outcome of more or less deterministic action (vide Section 1.3 for further discussions).

1.1.8. It would take quite some time before statistical studies form a regular branch of studies on language and literature. There

is little doubt that even now the statistical approach has yielded valuable results. The uses of style measures are admitted by many. Word-counts and counts on letters, phonemes etc., are also agreed to be valuable. Information theory has thrown a flood of light on the nature of language as a system of codes. The descriptive value of all these findings is certainly great. It is true that we do not have satisfactory probability schemes for many linguistic phenomena. It is for linguists and statisticians to work together in this direction.

1.1.9. This, the introductory chapter, attempts a bird's-eye-view survey of the previous researches, with emphasis on lines of research pursued in the present study. The object is to place the present researches against the background of previous ones. Unfortunately, for reasons of space, numerical results could not be quoted in detail. The interested reader is referred to Miller (1951, Chapters 4 - 7) and to Herdan (1956) for other surveys of the same field, and especially for numerical results.

1.1.10. The survey is carried out under three broad heads :
(i) Statistical studies on literary style (stylostatistics),
(ii) Studies on statistical properties of languages (statistical linguistics) and (iii) information-theoretic analyses of languages. This last category might come under category (ii), but deserves separate treatment, in virtue of the deeper methods employed in them. The three categories of researches are surveyed in Sections 1.2, 1.3 and 1.4. Section 1.5 gives a short account of some earlier work done in

India; these are only briefly mentioned in appropriate places in the three preceding sections.

1.2.1. Statistical Studies on Style : Style is an essentially statistical concept, being based on whole works rather than on isolated phrases or the like. When reading a work, the careful reader may form an overall impression about the style of the author, noting many different aspects of style at the same time, like word-length, sentence-length or size and composition of vocabulary. The statistical approach has to take up one aspect at a time, and to construct an objective measure of each aspect. So the statistician can cover only a small number of aspects in this way. However, subjective impressions are not sufficiently precise, nor fully retraceable or communicable; being subjective, they may be seriously misleading, having been coloured by a few striking words or phrases (vide Yule, 1944, Chap. 1). The statistical approach is free from these limitations.

1.2.2. Miller (1951, p.119-20) reports experiments by Allport et al / and Vernon / where a number of subjects wrote essays on the same themes at different periods. The essays written by the same subject could easily be grouped together by the judges, but the judges could not explain how they actually did it. It is obviously desirable to have objective indicators of individual style.

1.2.3. Statistical style measures are many and varied, and may be put to numerous uses. Some of the classical work in this field have been concerned with problems of disputed authorship or of determining

the dates or order of composition of different works by an author. They may also be used to obtain quantitative ideas about styles of different authors, in different fields of literature (e.g., fiction or journalism) and in different periods. Intercorrelations between different authors may also be studied.

1.2.4. Booth et al (1958, Chap.4) give a short account of the work on Platonic chronology (vide Lutoslawski,1897 , for a detailed study). The three dozen or so 'dialogues' attributed to Plato show diverse doctrines which are sometimes contradictory. There is thus a necessity of knowing the order of composition of these dialogues, so that it becomes clear as to which represent his early beliefs and which his mature doctrines. Very little internal or external evidence is available on this point.

1.2.5. This problem was discussed ever since the Renaissance, but little could be achieved till in 1867, Lewis Campbell introduced the stylistic method. He compared the technical vocabulary of the dialogues with that of the 'Laws', whose date was approximately known, and used the closeness of agreement as the indicator of date. The method was rediscovered in 1896 by W. Dittenberger who stressed, more correctly, the occurrence of certain everyday words. Since then a large number of persons have attacked the problem,some emphasising the more significant words, some the less significant parts of speech, others the rhythm of prose etc.; and the dialogues have been divided into three chronological groups with considerable confidence. It is known that, under

the influence of the rhetoric of Isocrates, Plato began to change his natural style when he was about 50 years of age, and the change was completed by the time he was sixty. This is why the stylistic method succeeded in dividing the dialogues into three clear groups. The sequence within the group could be determined only for the last group where the style conformed more and more closely to the rhetorical rules of Isocrates.[1/]

1.2.6. The chronology of Shakespeare's plays is another frequently discussed problem, but here nonstatistical evidence is by no means scanty. Some statistical methods employed by Shakespearean scholars are briefly mentioned by Bradley (1904, Appendix Note BB). Bradley admits the risk of incompetent uses of such tests, but is also convinced of their value when they are properly used. The most sensitive statistical tests are concerned with endings of speeches and lines. "It is practically certain", writes Bradley, "that Shakespeare made his verse progressively less formal, by making the speeches end more and more often within a line and not at the close of it; by making the sense overflow more and more often from one line to another; and at last, by sometimes placing at the end of a line a word on which scarcely any stress can be laid. The corresponding tests may be called the Speech-ending test, the Overflow test, and the Light and Weak Ending test". The percentage of cases of each type, it is

1/ Vide in this connection the paper by Cox and Brandwood (1959) which applied a variant of discriminant analysis for arranging chronologically the dialogues written between the "Republic" and the "Laws". The material used was the frequency distribution of sentences according to the lengths of the last five syllables.

believed, should move in a particular direction with time[1].

1.2.7. A paper by Yardi (1946) may be mentioned here. Yardi criticised the non-rigorous studies carried out on metrical data for Shakespeare's plays; and applied advanced statistical techniques like discriminant function analysis on data (collected by others) relating to the number of redundant final syllables, full split lines and unsplit lines with pauses /vide Nardan, 1964, Chap. 7.7/.

1.2.8. Such methods of determining the order of composition of an author's works rests on the assumption that the statistical measure chosen changes progressively with age. In case the measure chosen depends on other factors like the subject-matter or mood of the different works, the statistical approach may be seriously misleading and open to criticism. /Vide Herdan, 1956, Section 2.2, on Kitto's criticism of the method of dating Sophocles' plays based on the relative frequency of resolutions./

1.2.9. An obvious and important indicator of style is the word-length. This was pointed out by Augustus de Morgan, the mathematician, as early as 1851 (Lord, 1958). Word-length may be measured in phonemes or syllables or letters. In an article published in 1887, Mendenhall, a physicist, used the frequency polygon of the percentage distribution of words by length in letters as the "characteristic curve" of the author. It appeared that this curve was surprisingly stable for different works of a given author. This furnished a means

1/ Incidentally, vide Bradley (op.cit. p.327) for interesting figures on percentage of prose in Shakespeare's plays.

of identification of the author, at least by exclusion, which is not as sensitive as a finger-print test, as claimed by Mendenhall, but might at least serve as a kind of 'literary blood test', to use the phrase used by Brinegar (1963) in his recent work.

1.2.10. In his 1901 study, Mendenhall applied this technique to solve the controversy regarding the authorship of Shakespeare's plays. He counted the lengths of two million words from the works of Shakespeare, Bacon and their contemporaries. It was found that different authors possessed different characteristic curves, but the curve for Marlowe nearly coincided with that of Shakespeare. This went against the theory that Bacon was the real Shakespeare, much to the disappointment of the Baconian financing the work.

1.2.11. Mendenhall's works remained relatively unnoticed until Hoffman (1955) drew attention to them and marshalled them as evidence in favour of the theory that Christopher Marlowe was the real author of Shakespeare's plays. Since then these studies have been reviewed and partly reproduced by Williams (1956), and also discussed by Gibson (1962) and Brinegar (1963).

1.2.12. The limitations of such statistical techniques and the relative advantages of the traditional non-statistical approach are amply brought out by Gibson (1962). Gibson makes a critical survey of the four principal theories concerning the authorship of Shakespeare's works, each theory setting up a contemporary of the Stratford actor as the true author of these works. He makes a devastating criticism

of all these anti-Stratfordian theories and, in particular, of the statistical evidence supporting the Marlowe theory (op cit, pp.140- 7). Gibson was wrong in overemphasising certain points, for instance, the insensitiveness of word-length to individual style; he did not seem to realize that the characteristic curve gives a more detailed comparison between authors than does the simple average of word-length, so that complete agreement between two authors is really unlikely. But there is little doubt that the statistical evidence was far from sound, simply because Mendenhall and also Hoffman lacked essential knowledge of this literary field. For one thing, Mendenhall included among Marlowe's works some which are known to be only partly due to Marlowe. Many other factors make such evidence of dubious value. One may mention the unreliable nature of the texts of Elizabethan works that have come down to us, and the fact that collaboration and revision of plays by others was the custom of the Elizabethan times. Gibson also points out that Mendenhall should not have compared Bacon's "Advancement of Learning" and "History of Henry VII" with Shakespeare's dramas, mainly in verse : Books of different genre are apt to give different characteristic curves. This means that the Baconian theory cannot be rejected on Mendenhall's evidence. In general, such tests should be based on works of the same genre[1/].

1/ Gibson's work is non-statistical excepting where he criticises the statistical evidence for the Marlowe theory. But ideas of probability and statistical significance are implicit in the discussions on (i) cryptograms discovered from Shakespeare's works and on (ii) the parallelisms between Shakespeare's writing and those of the claimants.

1.2.13. In the paper already referred to above, Brinegar (1963) made a careful application of what is essentially Mendenhall's technique to the problem of authorship of certain letters frequently attributed to Mark Twain. Brinegar applied χ^2 and two-sample t-tests, with due reserve, for comparing the word-length distributions.

1.2.14. Word-length distributions for different languages and works may be seen in Elderton (1949), Fucks (1952, 1955), Herdan (1956) and Oettinger (1954), to mention only a few of the many investigations. They are of great interest for comparisons between languages, between different fields of writing and also between and within authors. Some numerical results may be quoted here. The Dewey count on modern English prose (Dewey, 1923) gives 1.43 as the average word-length measured in terms of syllables; the averages in terms of phonemes and letters are 3.65 and 4.38 respectively. The average word-length in Shakespeare is only about 1.3 syllables, over 75% of Shakespeare's words being monosyllabic. The average is only 1.2 syllables in the "Authorized Version of the Bible" in English; but about 1.6 in essays by Macaulay or Gibbon (say).

1.2.15. The distribution of word-length is close to the geometric form for Fitzgerald's "Rubaiyat" of Omar Khayyam, but such cases are extremely rare (Elderton, 1949). Fucks (1955) showed that if x is the number of syllables in the word, x-1 follows the Poisson Law, to a rough approximation in many languages, though not in Arabic. Herdan (1958a) fitted the lognormal distribution to some data on

on conversational English where word-length was measured in letters or phonemes [vide also Williams (1956)].

1.2.16. Another widely used style-indicator is the distribution of sentence-length. This appears to be more sensitive than the distribution of word-length (Williams, 1940, 1956) being more style-conditioned than language-conditioned (Herdan, 1964, Section 1.3). Sir William Petty used about 60 words per sentence, on an average, in his "Economic Writings"; "Essays" by Macaulay, on the other hand, used only 22 (Yule, 1938). Averages as low as 8 are found in simple dialogues in English. The range of variation is thus much wider than for word-length.

1.2.17. Many studies on sentence-length are summarised by Miller (1951, Chap. 6). In most cases, the length had been measured in words; but in one study by Gray and Leary(Miller, **ibid**) length was measured in syllables.

1.2.18. Yule (1938) used sentence-length distributions for throwing light on an age-old problem of disputed authorship. The work discussed is 'De Imitatione Christi', and the rival claimants are Thomas à Kempis and Jean Charlier de Gerson. As in all statistical work, precise definitions were needed, and Yule discussed the problems of defining 'the word' and 'the sentence'. He finally decided to repunctuate the texts, where necessary. Except in modern times, he argued, punctuation did not get the attention it deserves and was partly due to the compositor. Yule showed that the sentence-length distribution for the

'Imitatio' (average length : 16.2 words) was closer to the distribution for other works attributed to à Kempis (average : 17.9) than to the distribution for religious works attributed to Gerson (average: 23). So, it is much more likely that à Kempis was the true author of the "Imitatio" than that Gerson was.

1.2.19. In a similar manner, Yule (1938) showed that Sir William Petty could not probably have written the "Observations upon the Bills of Mortality" usually attributed to Graunt. Yule also studied the sentence-length in essays by a number of English writers. The averages ranged from 22 words in Macaulay's "Essays" to about 60 in the "Economic Writings" by Petty, as already stated. Yule was satisfied with the consistency of sentence-length within authors, but for most of the authors he really sampled from one and the same collection of essays.

1.2.20. Williams (1940) showed that the distribution of sentence-length was approximately lognormal in certain essays by each of three English authors. (This conclusion, like those in para 1.2.15 supra, was not supported by any goodness of fit test; but a visual examination of graphs seemed to be adequate.) So only two parameters sufficed to specify the distribution, viz., the mean and the standard deviation (abbreviated s.d.) of the underlying continuous logarithmic variate. Some between author variation was found in respect of the mean of logarithms, which indicates the average of sentence-length; also while Chesterton and Shaw used about the same number of words per sentence, they varied in respect of the s.d. of log (length), on

which depends the coefficient of variation of the distribution of sentence-length.[1/]

1.2.21 A whole class of style-indicators is based on vocabulary, its size and composition[2/]. The size of the vocabulary is the number of distinct words (word-types) used in the text[3/]. It cannot be satisfactorily estimated from a written text, since an unknown number of rare words 'at the disposal of the author'(i.e., "at risk") have not been actually used. The sampling theory of this problem has not been fully solved (Corbet et al, 1943; Yule, 1944, Chaps.5 & 7; Good, 1953; Good and Toulmin, 1956; Herdan, 1960). The number of word-types increases less than proportionately with the number of word-occurrences (word-tokens) sampled. The course of this increase is, in general, unknown, and it is not possible to estimate the limiting size of vocabulary as the number of word-tokens becomes indefinitely large, without making important simplifying assumptions. Vocabulary comparisons between different texts should therefore be based on counts of equal size;;but even if a series of such comparisons is made with sample counts of various sizes, one is not sure if one can extrapolate to estimate the limiting vocabulary ratio.

1/ A recent study on sentence-length in Kannada prose by Subba Rao (1960) has been described in Section 1.5.

2/ An individual has at least three vocabularies — one for talking, another for writing and the third for reading (vide Miller, 1951, Chap. 6, in this connection). An average Englishman has a speaking vocabulary of the order of 5000 words, a writing vocabulary of about 10,000 words — the nearly 30,000 vocabulary of 'Ulysses' by James Joyce is extremely unusual — and a recognition vocabulary of well over 100,000 words. We are mostly concerned with the writing vocabulary as reflected in written texts. Figures for some authors and works are given by Herdan (1956, Chap.1).

3/ One must decide whether one is interested in root words (i.e., lexical units) or particular words (i.e., inflected forms), whether proper names will be excluded or not in the counting, etc.

1.2.22. Yule (1944, Chapters 5 and 7) discussed the problems and fallacies of using vocabulary ratios, percentage of a special class of words in the list of word-types, the ratio of special vocabulary to total number of word-tokens etc. All these vary systematically with the size of the word count. Such ratios had been wrongly used in earlier researches on Plato and Chaucer. Herdan (1956, Section 2.5) discusses the same ideas and applies them to the problem of authorship of a recently discovered work which might be Chaucer's.

1.2.23. Vocabulary diversification has been measured by intervals between repetitions of the most frequent word, usually 'the'; but it is obviously better to use the type-token ratio (TTR), that is the ratio of the number of word-types to the number of word-tokens. This ratio, however, decreases from 1 to 0 in a complicated way as the count-size is increased. Comparisons of TTR between texts of unequal length involves the use of the inadequately known relation between the number of word-types and the number of word-tokens. The slope of the standard curve (vide para 1.3.13 infra) has also been suggested, but the slope is very often -1, and its usefulness is dubious.

1.2.24. Yule (1944) constructed a measure (K) of vocabulary diversity which should be more sensitive to style. He confined his attention to nouns, but suggested that verbs and adjectives might also be examined. He studied the frequency distribution of distinct nouns[1/] in different works by Macaulay, Bunyan, à Kempis, Gerson etc.,

1/ Yule grouped together the inflected forms of the same root, broadly speaking (op. cit, pp. 32-33).

according to the number of occurrences (x) (x= 1,2, 3, ...) in samples of texts. While Bunyan used the same noun repeatedly, Macaulay tried to vary the expression by avoiding repetitions of the same noun. Let f_x denote the number of nouns occurring x times (x=1,2, ...). The frequency distribution changes in a complex way as the count size $S_1 = \sum x f_x$ is increased; all ordinary measures like the mean or s.d. change systematically with count size. Yule (op. cit., Chap. 4) demonstrated that $K = 10^4 \frac{S_2 - S_1}{S_1^2}$ is a "characteristic" of the distribution which remains fairly stable in samples of varying size drawn from the same work(s)[1]. Here $S_2 = \sum x^2 f_x$.

1.2.25. Yule (op.cit, pp. 48-49) put forward a probability model where the process of writing a text is regarded as equivalent to sampling words successively from an urn, say, at random and with replacement, different words having different probabilities of selection $p_1, p_2, \ldots$ with $\sum p_i = 1$. The observed distribution $\{f_x\}$ is, approximately, a compound Poisson distribution with truncation at zero. If the total number of words "at risk" be denoted by W, then $\frac{KW}{10^4} - 1$ estimates the square of the coefficient of variation of the probabilities $p_1, p_2, \ldots$, which is obviously independent of the length of text[2]. The stability of K mentioned above indicated that this probability model might be fairly realistic, and Yule applied the model in

---

1/ The factor 10^4 is introduced merely to avoid small decimals.

2/ For other discussions on the "characteristic" see Good (1953) and Herdan (1958b, 1956, Chap. 3). Herdan (1956, Chap. 3) demonstrates the stability of K in some Russian word-counts; but his conclusion that K has been decreasing with time in English texts is not justified by his Table 12, which shows K for only 3 works in English, excluding those by à Kempis and Gerson wrongly included in this table.

later chapters in the discussion of the problem mentioned in para 1.2.22.

1.2.26. Yule (op.cit, Chap.9) used this type of measure for a second attack on the problem of authorship of the "Imitatio". It was found that the size and composition of the noun-vocabulary and the characteristic K for the occurrence-distribution of nouns in the "Imitatio" was closer to those for other works attributed to à Kempis than to those for the theological works by Gerson.

1.2.27. Yule (op. cit, Chap.10) also studied the joint distributions of nouns according to number of occurrences in samples from the "Imitatio", from the other works of à Kempis and from the theological works of Gerson, taking two of the three samples at a time. The correlation in usage of nouns was higher between the "Imitatio" and other works of à Kempis than between the "Imitatio" and the works of Gerson or between the works of à Kempis and the works of Gerson. Herdan (1956, Chap.4) studied such inter-correlations between noun-vocabularies of six authors in works concerned with political subjects.

1.2.28. As regards the composition of vocabulary, the relative frequencies of the parts of speech may vary characteristically between works. In particular, the occurrence ratio of verbs to adjectives is believed to be an indicator of style. It is presumably related to emotional factors, varying greatly between authors writing on the same topic. It can, however, change with the author's age and varies greatly between different types of writing (Miller, Chap. 6).

1.2.29. For pedagogical and other purposes, a statistical index of readability is obviously useful, and this may be included in the list of important style characteristics. Readability means ease in comprehension and not the artistic merit of the passage; it is reflected by the accuracy with which readers answers comprehension tests based on the passage.

1.2.30. Many factors can be suggested as having some influence on readability : the proportion of less frequent or unfamiliar words, the type-token ratio, word-length, sentence-length, frequency of personal references etc. Experiments have been carried out for studying the (multiple) correlation between such factors and the readability scores observed in tests of reading comprehension. Among the many resulting readability formulas, the most well-known is the one due to Rudolf Flesch given in his "The Art of Plain Talk" (Flesch, 1946). In a later paper (Flesch, 1948) Flesch proposed a measure with two components, both ordinarily ranging between 0 and 100:

(i) Reading Ease Score = 206.84 - 0.846 (average no. of syllables per 100 words) - 1.015 (av. sentence-length in words),

(ii) Human Interest Score = 3.635 (percentage of personal words) + 0.314 (percentage of personal sentences).

This approach removes some of the short-comings of the earlier formula and is also more convenient in practical application.

1.2.31. Flesch (1948) gives detailed instructions for counting words, syllables, sentences, personal words and personal sentences. Roughly speaking, personal words include all nouns with natural

gender and all pronouns except neuter pronouns. Personal sentences include those spoken, with quotation marks or otherwise, those addressed to the reader etc. When the text is long, the averages and percentages are to be estimated from "systematic" samples of the text, by taking every third paragraph, say, or every 10th page, and not from selected "typical" passages.

1.2.32. Flesch (1948) indicates readability scores for texts at different levels of readability and also the school grade of students corresponding to each level. He also shows how sentence-length or word-length can alone be used as indicator of readability. Thus, the average sentence-length should be 8 words or less for very easy English, 11 for easy, 14 for fairly easy, 17 for standard English, 21 for fairly difficult, 25 for difficult and 29 or more for very difficult English. Scientific English often employs an average of 30 words or so and literary English an average nearly 20.[1/] In a similar manner, Flesch gives readability levels by word-length alone : very easy -1.23 or less, easy - 1.31, fairly easy - 1.39, standard -1.47, fairly difficult - 1.55, difficult - 1.67 and very difficult - 1.92 or more.

1.2.33. Flesch (1946) discusses the concept of readability in a highly entertaining manner; reveals the true nature of unreadability so far as English is concerned; and gives practical rules to be followed by writers for writing readable English. According to him, Basic English and other attempts (like those of Thorndike) to

1/ Dewey (1923, Table 2 of 1950 edn.) presents some figures for modern English showing that the average sentence-length in the material covered was nearly 20 words per sentence.

simplify English put unnecessary emphasis on vocabulary control; readability does not depend so much on vocabulary size alone (Flesch, op. cit, Chaps. 20-21).

1.2.34. An elegant analysis on initial letters was carried out by Yule (1944, Chap. 8). The distribution of initial letters of distinct nouns was very different for works by Macaulay and by Bunyan. Yule showed that the difference between the two distributions is mainly due to the much larger proportion of Latin-Romance nouns (and a correspondingly lower proportion of old English-tutonic nouns) in Macaulay's works and the differences in distribution of initial letters between the two above-mentioned streams of English nouns. More generally, the relative frequencies of words in different etymological groups should give important indicators of style in many languages (vide Section 1.5 for an account of Chatterjee's investigation on Bengali).

1.2.35. Finally, some interesting findings on relative frequencies of punctuation marks are quoted by Miller (1951, Chap. 6). These depend partly on subject-matter etc., but partly also on the taste of the author. There was also some historical trend over the past few centuries.[1/]

1.2.36. Studies on lines mentioned above reveal important differences between languages. One may consider, for example, the distributions of word-length presented by Fucks (1955) for different languages; or the differences in sentence-length between English prose (Yule, 1938) and Kannada prose (Subba Rao, 1960). Within a given

1/ Vide Dewey (1923, Table 2 of 1950 edn.) for some data on modern English.

language, again, the style measures vary between broad fields of literature. Thus, according to Harris (1959), the average of word-length is between 1.5 and 1.6 syllables in prose dramas and letters by Goethe and Schiller written in the colloquial language, but between 1.8 and 1.9 in letters of good style, narratives etc. written by the same authors. (Note also the differences between prose and poetical works by Gray and Swinburne, shown in Elderton, 1949 .). It is not very clear, however, how far these measures may be taken as indicative of individual style, that is to say, how far the measures are constant within an author for works in the same broad field.

1.2.37. The impression that style measures like word-length are approximately constant within an author seems to prevail in many quarters (vide Fucks, 1955, pp. 154-155), but the evidence in support of this conclusion is not at all convincing. The studies on the chronology of Plato's or Shakespeare's works clearly show that style can vary with the author's age. It may be argued that word-length is sensibly constant within Shakespeare's plays; more generally, it may be pointed out that Mendenhall's first study satisfied him about the consistency within authors (vide paras 1.2.9 - 1.2.10 supra)[1/]. But possibly the consistency may be greater among dramas; also investigations on a few authors cannot lead to very general conclusions.

1/ It is not quite clear from the various accounts of Mendenhall's work whether Mendenhall really had reason to be satisfied on this point. Even Yule (1938) concluded that sentence-length is an indicator of style and applied it to a problem of disputed authorship without really examining within author differences in this respect.

Style may also vary with the subject-matter or mood of the works, within any given field of literature, and at any particular age of the author. This has been the experience of the present investigation on Bengali prose. ~~As stated by Williams (1940, p.361).~~ Literary Bengali underwent a rapid change during the last 100 years covered in this study. It is expected that in a slowly evolving language like English, within author differences would be less serious than in Bengali. /As stated by Williams (1940, p.361) Large scale trials are needed on this point. In any case, when applying statistical methods to problems of disputed authorship, one should not tacitly assume that within author variation is negligible without quantitative investigations into the matter.

1.3.1. Statistical Studies on Languages : The major line of investigation is to estimate the relative frequencies of verbal units — phonemes, letters, syllables and words. Relative frequencies of pairs and triplets of phonemes/letters have also been occasionally estimated. For letters, phonemes and their combinations, the relative frequencies seem to be fairly stable within a language, or rather the actual texts seem to be wellnigh random. This had led to Markov process models of speech samples or texts considered as sequences of letters or phonemes (vide Section 1.4).

1.3.2. Differences between works within a language are usually small, if not completely explainable by sampling error. Herdan (1955) rightly points out that this is partly due to the sound-sense correlation being weak in languages (Jespersen, 1922, Chap. XX). Herdan (1956,

Chaps. 5-6) also shows the importance of other factors like overlap in vocabulary, especially the grammar words. One may refer in this connection to Herdan (1956, Section 6.4) where Boldrini's study on Italian phonemes is discussed. Boldrini found that between author variation is almost negligible in poetry, but appreciable in political writings. Apparently, the distribution of phonemes contributed by lexical words depend appreciably on subject-matter.

1.3.3. Relative frequencies of phonemes have been estimated for many languages. Dewey (1923) gives the data on about 40 phonemes in modern English prose. Vowels and diphthongs cover about 38% of the phonemes and consonants the remaining 62%. Estimates for some other languages will be found in Herdan (1956, Chap.5) (vide also Zipf, 1949, Chap. 3). Herdan (**ibid**) gives some tables where the phonemes are grouped by organ of production. Whitney (1923) furnished the estimates for Sanskrit and Chatterjee (1926) for Bengali. Chatterjee's study is described in Section 1.5. Such estimates throw light on the structure of the languages. The distance between the phoneme distributions in two languages was measured by Forstemann by means of a statistical index which is defined as $\sum_i |p_i - p_i'|$, where p_i and p_i' are the percentage frequencies of the ith phoneme in the two languages (Herdan, 1956, Section 5.4). These indices were calculated separately for vowels and consonants. Zipf (1949, pp. 97-109) tried to explain the relative frequencies of different phonemes in terms of relative ease of articulation and auditory discrimination. Apart fro-

such theoretical interest, relative frequencies of phonemes or syllables are useful for evolving systems of shorthand (Dewey, 1950 edn., preface).

1.3.4. Relative frequencies of 1370 commonest English syllables were also estimated by Dewey (1923), separately for different positions within words. These accounted for 93% of the 143,000 syllables found in the 160,000 word-tokens. The total number of different syllables was 4400. French, Carter and Koenig (1930) studied 80,000 word-occurrences in 500 telephone conversations between businessmen. This revealed many interesting facts about composition of syllables; for example, syllables of the type CVC (consonant-vowel-consonant) formed 33.5% of syllables in spoken English, CV 21.8%, VC 20.3% etc.

1.3.5. Relative frequencies of letters are also available for many languages; for the data on English, vide Dewey (1923). Relative frequencies of pairs and triplets, called digrams and trigams, are available for English in Pratt (1942). A recent study by Bourne and Ford (1961) presents the relative frequencies of English letters, separately for each position within word, and also of digrams. These are valuable for cryptographic work (Hardan, 1956, Section 7.3), for evolving telegraphic codes, which should be shorter for more frequent letters, for information-theoretic analysis of languages (Shannon, 1951; also Section 1.4 below), for developing typewriter keyboards (Dewey, 1950 edn, preface), and also in connection with printing.

1.3.6. A common line of investigation is the word count. "As early as A.D. 900, the Talmudists were counting the words and ideas in the Torah in order to find out how many times each word appeared and how frequently each word appeared in an unusual way" (Miller, 1951, p. 88). The primary objective of word-counts is to determine the number of distinct words and also their individual frequencies; but a word-count can be put to a variety of uses.

1.3.7. Researchers have published the results of word-counts in different languages. A brief account of some English word-counts is given in Dewey (1923, rev. edn. 1950) and a detailed account may be seen from Fries and Traver (1940). Word-counts should ordinarily keep every derivative or inflected form (i.e., every particular word) separate, although for certain purposes one should group all variants (like 'large', 'larger', 'largest', 'largely') of the same root word or lexical unit (here 'large'). Many investigations prior to Dewey's followed an arbitrary mixture of the two approaches (Dewey, op. cit, preface). Many of them were also based on specialised types of material, like personal or business letters. The Dewey count of written English was based on sound principles and on carefully diversified material covering 100,000 words from modern English prose. In addition to relative frequencies of phonemes, syllables, letters and punctuation marks mentioned earlier, Dewey gave a list of the commonest 1027 particular words occurring more than 10 times in the count, and another list of 1132 root words occurring more than 10 times. The total number of

particular words was 10119 and the 1027 commonest among them accounted for 78634 occurrences. The 1132 root words covered 87380 occurrences out of 100,000.

1.3.8. Most word counts published only the commonest words; beyond the 500 or 1000 commonest words, in fact, the list depends critically on the type of material (Dewey, op. cit, p.18).

1.3.9. We must also mention "The Teacher's Handbook of 30,000 Words" by Thorndike and Lorge (1944), which "is based on counts covering 18 million English words from different types of material, including juvenile literature". French, Carter and Koenig's count has been mentioned in para 1.3.4 . They published a list of the 738 words (out of a total of 2240) appearing 5 times or more accounting for 76054 occurrences. Hanley (1937) has given an excellent "Word Index to Ulysses" by James Joyce. The appendix by Joos shows that there are 29899 different particular words among the total of 260,430 occurrences. This index also gives the page and line reference for each occurrence of each word occurring upto 24 times. Yule's counts on nouns in four works by Bunyan are presented as appendices to Yule (1944)[1/].

1.3.10. The uses of such word counts are many (Dewey, op. cit, preface). As already stated, they indicate the size and composition of vocabulary, but there are difficulties of extrapolating to estimate the limiting picture for indefinitely large counts (paras 1.2.21-24 supra). Word counts show the dominance of a fairly small number of

1/ See Section 1.5 for an account of some work on Bengali by Deb Chaudhury (1931) and Roy and Roy (1946).

common words, especially the grammar words; thus, the 50 commonest words comprise nearly 60% of all occurrences in spoken English and 40 to 45% of the occurrences in written English. The major use is in pedagogy, in the preparation of graded word-lists, broadly in descending order of frequency, for teaching pupils at different levels. Such lists are given by Thorndike and Lorge (1944); at any stage, the vocabulary learnt has the maximum utility for the given size. Word counts have also been useful for the evolution of Basic English (Ogden, 1934) where the emphasis is on vocabulary control. This is a simple adaptation of English proposed as an international language. It has only 850 words carefully chosen from English, excepting some additional technical words. Flesch (1946, Chaps. 20-21) criticises the word counters and also Ogden for having overemphasised the need of vocabulary control.

1.3.11. A word count may form the basis of many types of studies some of which may be carried out without first tabulating the frequencies of words. Thus, relative frequencies of phonemes, syllables or letters may be estimated, as in Dewey (1923); word-length distributions may be prepared; or one might examine the relative frequencies of words in different parts of speech or etymological groups. Miller (1951, p.89) states that the Kaeding count of nearly 11 million words in written German (Kaeding, 1897-8) shows that even in German 50% of the words are monosyllables and only 8.4% have 4 syllables or more.

1.3.12. Some word counts are called semantic counts. These study the frequencies of different words separately for each meaning of each word. They provide the basis for more functional dictionaries which show only the commonest meanings of the words. Thus Thorndike Senior Dictionary is based on the semantic count by Lorge (1937).

1.3.13. A celebrated result based on the word count is the rank-frequency relation studied extensively by Zipf (1949, Chaps. 2-4)[1/]. It was observed by Estoup in 1916 and by Condom in 1928 (vide Zipf, op. cit, References, note no. 4 on Chap. 2). The relation is this: If a complete word count is carried out on a sufficiently long text, the frequencies of different particular words tend to form a harmonic progression. If the frequency of the rth most frequent word is f_r, then $rf_r = c$ (constant), approximately, for all r. If f_r is plotted against r on double logarithmic scale, the curve resembles a straight line with a slope of -1 (the standard curve), except that horizontal steps of progressively increasing length appear at the highest ranks. (This latter is obviously because the frequencies are necessarily integral.) The relation is approximately equivalent to the statement that word-frequencies follow the Pareto type of distribution, the number of words having frequency x being inversely proportional to x^2 or $(x^2 - \frac{1}{2})$.

1.3.14. This rank-frequency relation has been observed in a wide variety of texts written in very different styles in different types

1/ Hence sometimes mentioned as the 'Zipf law'.

of languages in different periods of history. It also appears to be permissible, within limits, to pool word counts from different types of material without distorting the relation. Deviations from the 'standard curve' have also been observed (Zipf, op. cit, pp.121- 9). Thus, when the size of the word count is too small or too large, the slope may somewhat differ from -1. Actually, the slope tends to increase, though very slowly, with increase in the count size. Second, the curve may show a top-downward concavity near the dozen or so most frequent words, if the material is based on informal or colloquial speech or writing, where frequent words like the article 'the' are frequently omitted. Third, the slope may be less than 1, in the absolute sense, for holophrasic languages like Nootka or Plains Cree or present day Palestinian Hebrew. And so on. Nevertheless, the fit given by the Zipf law is often very good. The wide range of application of the 'law' has most probably some deeper cause and various probability models have been suggested for explaining this aspect of human behaviour, models leading, approximately, to the Zipf law of word frequencies.1/

1.3.15. Two such models were put forward by Mandelbrot (1953, 1954). The first makes the unrealistic assumption that actual languages are optimal (most economic) codes in a certain sense. The second assumes that the sequences of letters forms a Markov chain where one state is specialised to act as the space between words. This model gives .

1/ Yule (1944, p.55) says that the Zipf law is unlikely to be true, since it leads to an infinite variance, if not an infinite mean, of the distribution. This, however, is not as unreasonable as Yule supposed.

considerable insight and does not assume optimality of languages[1/]. It is, however, far from perfect, as would be evident from Miller, Newman and Friedman (1958) [vide also Miller (1957) and Miller and Newman (1958)]. That the Markovian structure cannot be a perfect model for human languages is shown by Chomsky (vide Miller, 1964, p.249).

1.3.16. Simon (1955) proposed a stochastic model of writing a text word by word which led to the Zipf law, approximately. His attempt to unify the Pareto law of income distribution, the Willis-Yule distribution in taxonomy (Yule, 1924) and the Zipf law for word frequencies, was really valuable from many points of view. But Mandelbrot (1959) showed the inadequacy of Simon's work, and a heated controversy ensued (Simon, 1960, 1961a, 1961b; Mandelbrot, 1961a, 1961b). Herdan (1961) also pointed out some defects of Simon's model.

1.3.17. Relatively unnoticed is the model due to Good (1957) where it is assumed that the effort in extracting the rth commonest word from memory is approximately proportional to log (r + a), where a is a constant, independent of r. Under this assumption the expected amount of information per unit of effort is maximum when word frequencies follow the Zipf law.

1.3.18. Incidentally, Herdan (1964, Chap. VII) argues that since the word-frequency distribution preserves its general form when different word counts are combined, the distribution must be of the compound Poisson type, as assumed by Yule (1944).

1/ Mandelbrot arrived at a generalised form of the Zipf distribution, viz., $p_r = P(r + \rho)^{-B}$ where p_r is the probability of the rth most frequent word and P, B and ρ are constants. This is sometimes called canonical form.

1.3.19. The Zipf law is not satisfied by words occurring after a specified word like 'of' or by words beginning sentences or by nouns (Miller, 1951, Chap. 4). The same negative conclusion is reached when content words and function words are examined separately (Miller, Newman and Friedman, 1958). So the law only provides the overall picture, ignoring finer points of language structure.

1.3.20. Zipf (op. cit, Chap. 4) studied some samples of speech of infants and children, of age ranging from 22 months to 7 years. The rank-frequency relation was closely observed even in such material, provided one includes the echolalia and the spontaneous acultural utterances (gibberish) which occurred in decreasing amounts with increasing age. Zipf utilised the findings for some discussion on the origin of speech. Miller (1951, Chap. 7) reports on these and other studies on verbal behaviour of children, the growth of vocabulary and complexity of sentences with increasing age.

1.3.21. We may now present the outlines of Zipf's other investigations on language (Zipf, op. cit, Chaps.2-3). He put forward the principle of least effort as an universal principle governing all spheres of human activities. Words were to him like a set of tools used for conveying meanings in order to achieve different objectives. These tools are therefore subject to forces which tend to minimise the expenditure of effort. General considerations led him to anticipate many 'economic' principles underlying linguistic phenomena, and he demonstrated the operation of many of these by means of data on a

variety of languages. Briefly speaking, Zipf concluded that the more frequent tools, say, words, would tend to be shorter, older and more versatile (by conveying a larger number of meanings) and would tend to enter more frequently into permutations (holophrases).

1.3.22. Zipf demonstrated (Zipf, op. cit, pp. 63-66) the widely known phenomenon that the more frequent words are shorter on the average, by using data on English and Latin where the regression of x on y has a negative slope. This was called the law of abbreviation (of frequent words). This means there is a negative correlation between lengths (x) and frequencies of occurrence (y) of different words found in a word count. Zipf states that this negative correlation has been found in a wide variety of languages. Formal and semantic changes are continually occurring in languages, acting on the whole in the direction of shortening the more frequent words or of increasing the frequencies of shorter words. Herdan (1956) mentions this negative correlation in some places of his work and presents some Tables (op. cit, pp. 74-75, pp. 140-143) showing the downward slope of the regression of y on x. Herdan (1958a) suggested the form $\bar{y}_x = ax^{-b}$ for the regression of y on x and showed that b is nearly 2.4 for some material on conversational English, with word-length measured in letters or phonemes.[1/] All these studies were concerned with the regressions of y on x or vice versa, but measures of correlation do not seem to have been used.

1/ Miller and Newman (1958) discussed this negative correlation in an attempt to explain the well known rank-frequency relation for words. See also Miller, Newman and Friedman (1958).

1.3.23. Zipf (op. cit, pp. 109-120) showed that as we consider rarer and rarer words of a language, we get a larger and larger proportion of the newer (nascent) words of the language. Also, within words of given frequency range, the longer words tend to be the newer words.

1.3.24. By utilising the Thorndike Senior Dictionary based on the semantic count by Lorge, Zipf (1949, pp. 27-31) showed that the number of different meanings m_r conveyed by the rth most frequent word in English tends to be proportional to the square root of its frequency f_r. (Zipf used the 20,000 most frequent words given in E. L. Thorndike (1932) : A Teachers' Word Book of 20,000 words, New York, Teachers' College, which was based on a 10 million word count.) This, the law-of-meaning distribution, supported Zipf's principle of economic versatility of tools.

1.3.25. In support of the principle of economical permutation, Zipf (op. cit, pp. 76-86) showed that in holophrasic languages like Nootka, the rank-frequency curve is of the standard shape after holophrases have been split into constituent morphemes or varimorphs, but the curve for holophrases and independent words is flatter. This indicates that the more frequent words enter more frequently into holophrases. In pp. 87-96 of the same work, Zipf shows that the r - f curve for morphemes is even steeper than that for words, which indicates that the more frequent morphemes combine more frequently into words. The r-f curve for morphemes is not linear on double-log scale.

1.3.26. Zipf regarded the r-f relation itself (**op. cit.** pp.19-22) as representing a balance between two (economic) forces : One, the force of unification tends to reduce the vocabulary to one word conveying all possible meanings (most convenient to the speaker), and the other the force of diversification, tending to increase the vocabulary to the point where each word has only one meaning.

1.3.27. Finally, in pp. 97-109 of the same work, Zipf shows certain economies of the phonemic systems of human languages. Of particular interest is his emphasis on the similarity of relative frequencies of consonantal phonemes in different languages. He suggested that these relative frequencies are governed by the relative ease in articulation and auditory discrimination. Thus, voiceless stops are generally more frequent than voiced ones, and short vowels than long vowels and diphthongs.

1.3.28. A number of probability problems of interest to the linguist were discussed in a remarkable paper presented before the Royal Statistical Society of London by Ross (1950). Some of Ross' views have been mentioned in Section 1.1.[1/] Among the seven probability problems discussed by him were (i) the problem of testing whether two languages

---

1/ Ross' criticism of stochastic models of language has already been referred to in para 1.1.7. Some of the discussants could not accept his point of view. Ross also asserted that in problems of disputed authorship, the traditional methods are all-important, those depending on unique and rare features, words, meanings and constructions. Where traditional methods are applicable, the statistical method is only supplementary; and where traditional methods are not available, statistical methods can never be conclusive.

are descended from a common parent, (ii) assessing evidence for close relationship between two related languages, and (iii) arranging the books of the Rgveda in chronological order on the evidence of the frequencies of features which, rare in early Sanskrit, became increasingly frequent with time. Ross suggested fairly advanced statistical methods for solving some of these, and the discussants suggested various alternative and supplementary solutions. Wake, in this connection, tried probit-analysis methods on Yule's data on frequencies of nouns.

1.3.29. Herdan (1964, Chap. IX) describes some exploratory work in a line indicated by Ross. For each pair formed from seven Indo-European languages, he considers a 2 x 2 table, the cell-frequencies denoting the number of all Proto-Indo-European (PIE) roots possessed by both or either or none of the two languages. Herdan applied Spearman's two-factor analysis to measures of association obtained from these tables, and estimated saturation of each language with the PIE as the common factor. However, presence of other factors like geographical location are also indicated.

1.3.30. Fucks (1954) started a new line of investigation. When the words in a text are replaced by their lengths and these lengths are read in the natural reading order, one gets what may be called a word-length series. Fucks considered the randomness of such series for a number of texts in German and for one in English ("Othello" by Shakespeare). Lengths of consecutive words seemed to be roughly independent in the statistical sense; and the first order auto-

correlation coefficient was very close to zero. There is no such study on any series of sentence-lengths. However, Yule (1938, p. 371) remarked that this correlation is expected to be positive and appreciable for some authors.

1.3.31. An important problem is to study the lengths of intervals between successive occurrences of the same word. Zipf (1949, pp. 40-54) reports on a number of researches mostly relating to rare words in "Ulysses", using the page references for the occurrences given in the Hanley "Index" referred to in para 1. 3. 9 . By using faulty statistical methods (vide Chap. 11 of this work) Zipf reached the following conclusion : The number of intervals n_I of length I (measured in pages) between repetitions of words occurring f times each in "Ulysses" was approximately given by $n_I^p \, I = C$, where C and p are constants, with p usually lying between 1 and 1.25.[1/] Secondly, the successive interval lengths showed the same kind of distribution, that is, intervals of different sizes seemed to be evenly distributed over the entire text. Zipf also presented some data on low-frequency words in an Old English epic (Beowulf) and on Homer's Iliad, measuring intervals in lines. Linearity of the relation on double-log scale was not quite satisfactory in these cases, but again, successive interval lengths seemed to be equally distributed.

1/ Zipf fitted this relation which is linear on double-log scale by considering only the range I = 1 through 21 (1 through 50 in one case). But the linearity does not hold for larger values of I. This is a major defect in Zipf's procedure.

1.3.32. Herdan (1956, Section 6.7) reports a small piece of investigation on a high-frequency Russian grammar form ( ) showing that the distribution of intervals I, measured in words, is approximately geometric, i.e., $p_I = p^I (1 - p)$, (I=0, 1, 2, ...), where p is a constant, and p_I the proportion of intervals of length I. (Actually, Herdan used the exponential approximation to the distribution function of I.) This model has a beautifully simple probability interpretation.

1.3.33. Yngve (1956) suggested a statistical procedure for studying the frequency distributions of intervals between occurrences of any two specified words and for examining whether the two words occur independently of each other. Such studies would throw light on the constraints operating between different words in a language.

1.3.34. Such investigations deserve great deal of attention for they throw light on the fundamental question of the applicability of probability models to texts or samples of speech considered as sequences of words. The model used by Yule (1944) (vide para 1.2.25 supra) was severely criticised by a linguist, Ross, who asserted that writing is an action where deliberate choice predominates and probability models like that of Yule are totally inapplicable. Yule's model may, admittedly, be an over-simplification of reality, but Ross' objection seems to be based on a misconception of the scope of probability models. It may be recalled that the distribution of German bombs over the area of London was found to follow the Poisson law to a very high degree of approximation (Feller, 1957, Chap. VI, Sec. 7). It is quite possible

that the outcomes of apparently deterministic action do in fact obey deep, unsuspected, probability laws. Indeed, the Zipf law, already mentioned, is one such law. Information is still meagre, however, for any definite conclusion in this matter.

1.3.35. Yule (1944, Chap. 7) himself had produced a strand of evidence against the deterministic view of the use of words. He examined the counts on nouns in four essays by Macaulay and also in four works by Bunyan. In addition, the sample from Macaulay's essay on Bacon was split into four subsamples. In each case, the nouns were sorted according to whether they occurred in one, two, three or all four of the four samples or subsamples. The frequency distributions of nouns over these classes showed remarkable regularities and agreed, broadly, with what could be expected if the word-tokens appearing in the four samples (or subsamples) taken together were randomly partitioned among the four samples. The probability formulae of this model resemble the well-known Bose-Einstein statistics of theoretical physics, which therefore, according to Herdan (1964, Part V) represents a fundamental model for linguistic phenomena. It is obviously necessary to examine these ideas in more extensive material.[1/]

1.3.36. We mention now certain studies on the relative frequencies of different parts of speech (French, Carter and Koenig, 1930; Herdan 1956, Section 6.6) separately among word-types and word-tokens. The

---

1/ The nature of laws in the social sciences was discussed by Kendall (1961), who said that "not only can choice mimic chance, but chance can mimic choice" (Ibid, p.12).

words of the minor parts of speech, sometimes called grammar words (or function words) form a small proportion among word-types but about 50% or more among the word-tokens (in English). Miller (1951, Chap. 6) discusses the verb-adjective ratio as a statistical indicator of individual style (vide para 1.2.28 supra). The ratio may change with the age of the author. It is also very different in different types of texts — 9:1 in drama, 3:1 in fiction, 1 to 1.5:1 in theses and scientific writing etc. etc. It is also higher for spoken language than in written texts.

1.3.36. There have been several studies on the relative frequencies of words falling in different etymological groups. One was carried out by Zipf (1949, pp. 109-120) in connection with his study on the age-frequency correlation of English words (vide para 1.3.23). Another by Yule has already been mentioned in para 1.2.34 . Some others on Bengali, carried out by Chatterjee, will be described in Section 1.5.

1.3.37. Some studies of a miscellaneous nature may be finally mentioned, viz., one on the number and arrangement of spondees and dactyls in the Latin hexameter (Herdan, 1956, Chap. 5); another on the caesurae in early Greek hexameter (op. cit, Section 6.8); and a third on the distribution of lengths of runs of grammar forms and lexical forms in different languages (op. cit, Section 6.5).

1.3.38. So far as the present author is aware, probability sampling has not been systematically used for studies on languages or literary

style. Dewey (1923), Yule (1938, 1944), Williams (1940), Elderton (1949) — to mention only a few — mostly used various types of subjectively chosen samples. Admittedly, the problem is not very important when one is interested in relative frequencies of phonemes or letters (say) (vide paras 1.3.1—2). Also, Herdan (1956, Chap. 3) reports on some sampling studies by Epstein suggesting that, even for word counts, certain types of non-probabilistic samples (called 'spread samples' — see below) are effectively random. But surely if modern methods of statistical inference are to be applied, one should be more rigorous and cautious in one's approach and should not generalise from the few studies, if any, on the suitability of non-probabilistic samples.

1.3.39. Examples of probability sampling are extremely few. Yule (1938) tried certain methods of probability sampling of sentences, but his work was mostly based on non-probabilistic samples. In pp. 381- 3 of the paper, he discussed the problem of probability sampling, but apparently could not find any convenient, yet unbiased method. The subjective method permitted him to exclude non-typical matter from the sample when carrying out statistical tests of authorship. (But this could be done even in a probability sample !) For this and other reasons, he preferred 'the method of selected passages of considerable length' spread more or less uniformly over chosen works (spread sampling). He confessed that rigorous statistical inference was impossible. Yule (1944, Chap. 3) followed the same type of procedure for his now

classical studies on nouns. Apparently Yule could not think of probability samples other than unrestricted random samples. Here also Yule realised that he could not solve the question of sampling error of the characteristic K.

1.3.40. Since unrestricted random sampling would, in general, be too laborious, probability samples with some element of clustering seem to be the obvious choice. This was used to some extent by Yule (1938), as already stated. Herdan (1956, Chap. 4) selected pages at random for his sample counts on nouns in political writings of six authors. But again, while probability sampling was done in these cases, no attempt was made to examine the sampling errors. It is true that the problem is complicated, but Yule got very near to the practical solution in his 1938 paper. For each work, Yule divided his (non-probabilistic) sample of sentences into two or more "subsamples", and judged the reliability of the estimates by agreement between subsample estimates. This is analogous[1/] to the technique of interpenetrating networks of subsamples (IPNS) due to Mahalanobis (1946).

1.3.41. Quite often, the non-probabilistic method of selection resembled systematic sampling. Thus, Williams (1940) selected the first 30 sentences from each of the first 20 chapters of a work. For estimating measures of readability Flesch (1948) recommends going by a strictly numerical rule, say, every 3rd page or every 10th paragraph. Subba Rao (1960) selected all sentences appearing on about 50 pages (from each work) chosen in the systematic manner. It is quite likely

1/ Yule's subsamples were not interpenetrating, but based on different chunks of the work/s.

that these methods give more accurate estimates than unrestricted random sampling, but the trouble is there is no method of assessing the accuracy of nonprobabilistic samples (Yule, 1944, pp. 40-41). Williams (1940) calculated standard errors and Subba Rao (1960) applied the Kolmogorov-Smirnov tests and other techniques, assuming unrestricted random sampling, but it is obviously necessary to justify these assumptions, since the population is not effectively in a random order (Yule, 1938).[1/]

1.3.42. One of the major objectives of the present study is to show that for statistical studies on word-length, sentence-length, and the like, probability sampling can be readily and conveniently used. It is not claimed that the methods used are optimal from the joint consideration of cost and accuracy, but the path has been cleared for a judicious choice of sampling design. Another object has been to study how far non-probabilistic samples of the 'systematic' type can be used as reasonable approximations to probability samples. The question of sampling error has been constantly kept in view. In most cases, the inference was not based on standard errors calculated in the detailed way : The technique of independent and interpenetrating subsamples (IPNS) was found to be extremely serviceable in this respect.

1/ Fucks (1952, 1954) seems to have used complete enumeration, but does not state the fact explicitly, not does he quote any sample size. Even if complete enumeration had been carried out, the count size should have been given to give some idea of reliability.

1.4.1. Information-theoretic Studies on Languages : Information theory, one of the youngest branches of probability theory, was really born with the appearance of a now classical paper by Claude E. Shannon (1948). Shannon's treatment was intuitive and not fully rigorous : he was concerned with quick results for practical applications. Rigorous proofs of some of his theorems were given subsequently. [McMillan, 1953; Khinchin, 1957 ; see also Goldman's text book (1954) on information theory, and Barnard's (1951) paper on some basic theorems.] This theory is concerned with a general theoretical model for transmission of information of various kinds.

1.4.2. Information-theoretic analysis of actual languages was started by Shannon in the above-mentioned classical paper itself; but the methodology was really developed in Shannon (1951), where the results of an analysis of written English were presented. Since then other studies have appeared in this field. A brief account of some of these will be given below. In general, these studies are based on the theory of discrete communication on a noiseless channel.[1/] The relevant theory will also be summarised below, although in a somewhat non-rigorous manner.

1.4.3. Shannon (1948) indicated that messages in actual languages like English might be regarded as realizations of Markov processes of

---

1/ The study by Herdan (1956, Chap. 11) reported in paras 1.4.27—29 below uses the theory for a noisy channel, but information-theoretic analysis is not really essential there.

appropriate order, with fixed transition probabilities (vide Feller, 1957 , for an account of Markov processes). The letters (or phonemes) and the space are the 'states' of the process. The source of information behaves like a stochastic process and generates the message, a sequence of states, symbol by symbol, obeying the constraints imposed by the transition probabilities implicit in the digram frequencies, trigram frequencies etc., of the language. Shannon also assumed that the process is 'ergodic', which holds under certain weak conditions. This implies that any sufficiently long sequence generated by the process will, with probability tending to 1, have the same relative frequencies of letters, digrams etc. McMillan (1953) showed that the process need not be of the Markov type; the property of ergodicity would suffice for most of the theory.[1/] The stability of relative frequencies of phonemes, letters and their combinations (Herdan, 1956, Chapters 5-6) indicates that this assumption is fairly realistic for languages.

1.4.4. In general, for any ergodic process, sufficiently long sequences of N symbols (i.e., letters and spaces) fall into two classes. One class has nearly 2^{NH} sequences (where H is a constant), each with probability nearly 2^{-NH}, so that the total probability of this class of sequences is nearly 1. This class comprises all the sequences having relative frequencies of letters, digrams etc., close to the corresponding probabilities. The other class comprises all other

1/ Chomsky has shown (Miller, 1964, p. 249) that the Markovian structure is not completely adequate for natural languages.

possible sequences, with the relative frequencies not close to the corresponding probabilities. This class would have a total probability arbitrarily close to zero.

1.4.5. For practical purposes, the sequences of the second class may be ignored, and so the source of information can generate any one of 2^{NH} equiprobable messages with N symbols each. This describes a state of uncertainty and we want a measure of the amount of uncertainty. Sequences of NH binary digits (0 and 1) would suffice to encode all these 2^{NH} sequences. In a certain sense, therefore, a sequence of N letters/spaces generated by the source of information has the same amount of uncertainty (i.e., can show the same variety) as a sequence of NH binary digits or 'bits'. Since any realization of the stochastic process removes the uncertainty completely, we say that amount of information given by the observed sequence is equal to the amount of uncertainty before the realization. We, therefore, say each message carries NH bits of information, that is H bits of information per symbol.[1/] The method of estimating H will be given below. This H is called the entropy of the stochastic process, since it is analogous to the entropy in statistical mechanics.[2/] It has a number of elegant

---

1/ Note that we are concerned here with the coding or communication aspect of information and not with its semantic aspect.

2/ Thus, when successive symbols are independent, but the probabilities of different symbols are unequal, then $H = -\sum_i p_i \log_2 p_i$, where p_i ($p_i \geq 0$) is the probability of the ith symbol for $i = 1, 2, \ldots n$, ($\sum p_i = 1$). For a simple stationary Markov Chain, with transition probabilities denoted by p_{ij} $(i, j = 1, 2, \ldots \ldots \ldots n)$, the expression becomes $H = -\sum_{ij} p_i p_{ij} \log_2 p_{ij}$, where p_i has the same meaning as before.

and intuitively desirable properties which make it suitable as a measure of information or uncertainty (Shannon, 1948; Khinchin, 1957, Part I).

1.4.6. Suppose the process generating the message is such that all the n states (27 for English) are equiprobable and the successive symbols are independent. Consider sequences of N symbols. All the n^N possible sequences are equally likely here and the second class of low-probability sequences described above becomes empty. Since $n^N = (2^{\log_2 n})^N$, encoding these sequences of length N in the binary form would require $N \log_2 n$ binary digits, that is, $\log_2 n$ bits per symbol of the original message. We say that each symbol of the original message carries $\log_2 n$ bits of information. This is the situation where H reaches its maximum $\log_2 n$ (= H' say) and the effective number 2^{NH} of messages of length N, its maximum value $n^N = 2^{N \log_2 n} = 2^{NH'}$.

1.4.7. In other cases, the unequal frequencies of different states and the dependence between successive symbols arising from constraints due to spelling, grammar, syntax and idiom, e.g., the restriction that 'Q' must be followed by 'U' in English makes the effective number of messages 2^{NH} less than the permissible number $n^N = 2^{NH'}$, so that H is less than H'. The ratio H/H' is called the relative entropy, and its complement $1 - \frac{H}{H'}$, the redundancy of the message or its generating process. Now $2^{NH} = n^{NH/H'}$. Therefore, if the n states were used with equal probability and if the successive symbols were independently filled in, the actual message of N symbols could be compressed into NH/H' symbols

only. The actual message is therefore unduly long or redundant. It could be compressed to H/H' times its present length, if the n symbols were used in the most efficient way.

1.4.8. Redundancy due to statistical constraints in the use of symbols is not, however, an unmixed evil. That missing letters may be supplied or doubtful ones corrected is due to the redundancy of the language, for one uses in such situations the correlations or constraints between neighbouring letters which give rise to redundancy.

1.4.9. We come now to the actual estimation of entropy H. (The quantity H' may easily be obtained as $\log_2 n$.) Shannon (1948) showed that H may be determined by limiting operations directly from statistics of observed messages. Let $p(B_i, S_j)$ be the probability of a block of K-1 symbols B_i followed by the symbol S_j. Let $p_{B_i}(S_j) = \frac{p(B_i, S_j)}{p(B_i)}$ be the conditional probability of the symbol S_j when the preceding block of K-1 symbols is known to be B_i. Consider the conditional entropy of the Kth symbol when K-1 preceding symbols are known

$$F_K = -\sum_{i,j} p(B_i, S_j) \log_2 p_{B_i}(S_j)$$

where the sum is over all blocks B_i of K-1 symbols and over all symbols S_j. This F_k is called the K-gram entropy of the process. It is a monotone decreasing function of K, and $\lim_{k \to \infty} F_k = H$.

1.4.10. Shannon (1948) reported that the redundancy of ordinary English is roughly 50%, considering statistical structure of the

language upto groups of eight letters. The analysis was carried further and reported in detail in Shannon (1951). By using tables for relative frequencies of letters, digrams and trigrams, Shannon obtained the following estimates for English, ignoring the space symbol 1/:

$F_o = \log_2 26 = 4.70$, $F_1 = 4.14$, $F_2 = 3.56$ and $F_3 = 3.3$.

By exploiting the Zipf law of word-frequencies, a rough estimate of F_{word} was obtained as 2.62. Extrapolation gave $F_8 \sim 2.3$.

1.4.11. For $K > 3$, tables of K-gram frequencies are not available; and obviously the direct calculation of F_K for K = 100 or 1000 (say) is out of question. Shannon therefore devised a guessing experiment for estimating F_K for given values of K, large or small. This utilises the fact that speakers of the language possess implicitly an enormous knowledge of the statistics of the language. Samples of English text, 15 letters in length, were given to the subject, and the subject guessed the text, letter by letter, for each sample. (The subject used various tables for letter frequencies, digram frequencies etc., but these are not very essential.) Shannon obtained the frequency distributions of number of 'educated guesses' required to predict the correct Kth letter when preceding k-1 letters are known (k = 1,2,..., 15). A similar test was carried out where 100 preceding letters were known to the subject.

1/ If space be regarded as the 27th state, the value of F_N would be nearly $\frac{4.5}{5.5} = 0.818$ times the value when space is ignored, for large values of N, since an average English word has nearly 4.5 letters, and the space symbol is completely redundant.

1.4.12. Shannon developed some properties of an ideal predictor and used these to set up upper and lower bounds for F_K for K = 1,2, ...,15 and 101, from the above-mentioned frequency distributions, assuming that the subjects chosen were nearly ideal predictors. For K = 15, the bounds were 2.1 and 1.2; and for K = 101, 1.3 and 0.6. This meant that in ordinary literary English — Shannon's sample texts were from "Jefferson the Virginian" by Dumas Malone — the constraints upto 100 letters may reduce the entropy to about 1 bit per letter, redundancy being of the order of 75%.[1/] A 27-character alphabet (26 letters plus a space) could encode exactly the same information with about one-fourth as many characters of all character-sequences were possible and equiprobable.

1.4.13. Burton and Licklider (1955) observed that Shannon's estimates showed considerable increase in the relative redundancy between K = 15 and K = 101, the two highest values included in the experiment. It was not clear therefore how far the estimate for K = 101 could be taken as an estimate of the limiting value. They carried out some experiments on Shannonian lines using a larger sample of texts and subjects. The higher values of K:-1 were 16, 32, 64, 128 and (approxi-

---

1/ Shannon recognised that entropy may vary appreciably with the type of text; thus, newspaper writing, scientific writing and poetry would give higher values. Shannon said that the two extremes of redundancy in English prose might be represented by Basic English (Ogden, 1934) and by James Joyce's "Finnegan's Wake". Basic English is limited in vocabulary and a passage expands when translated into Basic English; Joyce's writing, on the other hand, achieves a great deal of economy, by employing a very large vocabulary. Quantitative studies in this direction would be extremely interesting.

...ly) 1000. The estimates seemed to level off at about K-1 =32, so ... the relative redundancy of printed English might be taken as nearly 75%, as indicated by Shannon.[1]

1.4.14. The Shannonian guessing game has several limitations; for example, the memory of the subject is unduly contaminated by what has gone just before. More objective procedures of estimating entropy were suggested by Newman and Gerstman (1952), and their intuitive ideas were elaborately examined by McGill (1954) and Garner (1958). For English texts the method gave results similar to Shannon's (Newman and Gerstman, 1952).

1.4.15. The method is based on measures of association between pairs of letters at varying distances from each other in a text. Denote by T(1, i) the conditional information in the ith letter that is dependent on the choice of the first letter. Then

$$T(1,i) = 2H(1) - H(1, i),$$

where H(1) is the entropy of the first (or any other) letter and H(i, i) that of the joint distribution of 1st and ith letters. Ignoring interactions of second and higher orders, the K-gram entropy is given by

$$F_K = H(1) - T(1,2) - T(1,3) - \ldots - T(1,K).$$

Garner's (1958) experiments indicate that including the second order interactions is hardly worthwhile. The method is especially useful for comparative studies covering several languages : Shannon's method would not be adequate for this, since no subject knows two languages equally well.

1/ Herdan (1964, p.255) states that German has 1.3 bits of information per letter, so that redundancy = $1 - \frac{1.3}{\log_2 32} = 0.74$.

1.4.16. Newman and Waugh (1960) applied this method for a comparative study of redundancies of three languages — Samoan, a Polynesian language, with a 16-letter alphabet, English with 26-letters and pre-1917 Russian with 35. They used identical passages from the Bible in these three languages, because, as shown by them for English, the estimates of entropy increase with the difficulty of the texts, as anticipated by Shannon (1951) (vide footnote to para 1.4.12 supra), and by Burton and Licklider (1955). While H(1) is larger for the language having a bigger alphabet, the values of F_{12} are nearly equal, 2.136, 2.397 and 2.395 respectively. These estimates are much higher than the true values of entropy, since constraints extending beyond 12 letters have not been considered. Some observations are also made on the usefulness of studies on comparative lengths of identical texts in different languages. For languages with larger alphabets messages were found to be shorter, but the differences were generally small, as the constraints were also stronger for languages with larger alphabets.

1.4.17. Bell (1953, 2nd edn, pp.126-29) used a different technique for getting a rough estimate of the 'internal information' of English words, that is, of the reduction of entropy due to constraints in the language. He estimated the number n_x of English words of x letters ($x = 1,2,3,\ldots$) from the Concise Oxford Dictionary, and compared these with the corresponding numbers N_x of possible combinations of x letters drawn from a 26-letter alphabet, without regard to pronounceability. The amount of internal information in a x-letter English word was

taken as $\log_2 (N_x/n_x)$ bits, which comes to $\frac{1}{x} \log_2 (N_x/n_x)$ bits per letter. The values for different x were then averaged to get the internal information of 2.1 bits per letter for all English words; the weights used were the relative frequencies of words of different lengths.[1/]

1.4.18. Bell and Ross (1955) followed the same line for a study on Welsh words, where the sound-symbol correspondence is not poor as in English. However, since no word-length distribution happened to be available, the internal information figures for different x could not be averaged as done for English by Bell (1953).

1.4.19. A recent study by Siromoney on the entropy of Tamil prose is described in Section 1.5.

1.4.20. Miller and Friedman (1957) pointed out that while theoretically English texts might be shortened to a quarter of their present length, the Fano-Shannon method of optimal coding (Shannon, 1948) has grave practical disadvantages. Blocks of K characters ($K \geq 32$) must be coded as a unit; also the coded message bears no resemblance to the original and has to be decoded before reading. The whole process is

---

1/ This is not comparable with the results of Shannon (1951) and others, which take account of long-range constraints extending over more than one word. Bell was satisfied by the agreement with Shannon's estimate of 2.3 bits per letter; but this was only a provisional estimate considering combinations of eight letters. Vide para 1.4.10 *supra*. Bell's method is also open to several criticisms. For one thing, the relative frequencies of words of different lengths were estimated from the 750 most frequent words in Dewey's (1923) list. This would surely give excessive weightage to shorter words in the language. Second, the estimates of n_x depend critically on the dictionary used.

slow and expensive. Miller and Friedman studied various methods of abbreviating English i.e., eliminating redundancy, without modifying basic rules of English orthography, e.g., by deleting letters at random. This was part of a detailed study on the abilities of human operators to correct various types of mutilations of written English texts. With superior persons and unlimited time, it is possible to achieve 50% compression, either by omitting alternate characters or by omitting all the vowels and the space between words. These lead to a lower bound of 60% for the redundancy of printed English. The paper contains a large number of information-theoretic results of considerable interest.

1.4.21. One general observation may perhaps be recorded here. While the concept of entropy (H) of a text/speech-sample is quite meaningful, the concept of relative entropy or redundancy is not. The maximum value of H, denoted H', corresponds to a situation where all possible combinations of letters are equiprobable. But many of these combinations cannot be pronounced. So the value of H' is put too high, and hence also the estimates of redundancy.

1.4.22. Fucks (1952) studied the word-length distributions for works in different languages and proposed $H = -\sum p_i \log p_i$ (where p_i = proportion of i-syllabled words) as an indicator of individual style, and also of the carrying structure of the language. Fucks did not demonstrate that within author variation in H was negligible. He did not also notice the fundamental defect of this entropy measure, viz. that H does not change if the p_i-values are permuted among themselves

or if all i-values are changed by the addition of a constant.[1] Nevertheless, Fucks (1952, also 1955) made a more reasonable use of the entropy measure than Herdan (1953, 1956, Chap. 10). Herdan's work seems to be based on a serious misconception of the 'fundamental theorem' of the theory of communication.

1.4.23. Given the proportions p_1, p_2, ..., p_r..of words having one, two, ..., r, ... syllables (say), Herdan calculates $H = -\sum p_i \log p_i$ and says that $H \leqslant \bar{x} = \sum_x x p_x$ according to the 'fundamental theorem', Herdan (1956, p. 172). This inequality can be proved by Lagrange's method of undetermined multipliers — the equality holds if and only if $p_r = \frac{1}{2^r}$ $(r = 1,2,\ldots)$ — and has very little to do with the 'fundamental theorem'. Since the lengths of successive words are approximately independent in the statistical sense (Fucks, 1954; present work, Chap.5), H does indicate the variety or information present in the word-length series obtained from the text. But it has nothing to do with the efficiency of the original text, and Herdan's attempts to study the relative efficiencies of languages (and also Pitman's Shorthand) by using the relative magnitudes of $\bar{x}$, H and H' were all fundamentally unsound.

1.4.24. We may finally mention certain studies on relative efficiencies or economies of different languages. Jespersen (1922, p.324) stresses that languages should be assessed mainly by their ability

---

1/ Similar comments may be levelled against the 'spur' $S = \sum \frac{1}{p_i}$ introduced by Fucks (1952). Since S is sensitive to the presence of a few large values of i, Fucks sometimes pooled the tail classes in a subjective manner. The same may be said about $H' = \log r$ (where r is the number of length-classes of words) used by Herdan (1953, 1956, Chap. 10).

to express the greatest amount of meanings with the simplest mechanisms. There is, however, no method of measuring either expressiveness or effort exactly. Jespersen (op. cit, Chaps. XVII-XVIII) demonstrated the progressive nature of evolution of languages and produced some statistical data indicating that the same idea can be conveyed by shorter messages in the more modern languages. The lengths in syllables of the Gospel of St. Mathew were approximately as follows in ten different languages[1]: Greek -39000, Latin - 37000, Swedish -35000, German -33000, Anglo-Saxon - 34000, French - 33000, Danish -32500, Gothic -31100, English -29000 and Chinese -17000. These figures support Jespersen's conclusion[2] (see also Miller, 1951, pp.114-5), but he admits that syllables in different languages are not strictly comparable. "The most rational measure of length would be the number of distinct (not sounds, but) articulations of separate speech organs .....". This brings us to energetics of speech. The techniques mentioned in Miller (op. cit, Chap. 2) should prove helpful in this matter.

1.4.25. In the paper already referred to, Newman and Waugh (1960) give lengths in letters of a section of the Bible in five languages : Russian -7115, French -7829, German -8062, English -7960 and Samoan-7656.

1/ Some later figures due to Baker (1950) are included here.

2/ Jespersen shows that modern languages use syntactical rules of word order to replace the system of inflections of older languages. This has led to a decay of inflectional affixes and has increased the percentage of monosyllables in the modern languages.

1.4.26. The value of such comparisons depends on the quality of translation, which is a noisy process. Also, as Jespersen (op. cit, p.330) pointed out, translations naturally tend to be more longwinded and verbose than the original. So when comparing English and German, say, one should use careful translations, of some English texts into German and some German texts into English.

1.4.27. We may note here an extremely interesting piece of study carried out by Herdan (1955, also 1956, Chap. 11). Herdan considered lengths of words in several passages in each of four European languages along with lengths of corresponding words in the English translations of these passages. Sometimes word-groups representing single ideas had to be considered, instead of words, because of lack of one-to-one correspondence words in the original and the translation. Herdan prepared a two-way distribution for each passage showing the joint distribution of lengths of words or word-groups in the original and in the translation. The distributions seemed to be reasonably stable for a given pair of languages, and showed marked positive correlation, partly because the translation equivalents are sometimes structurally related and partly because the same concept or idea has similar frequencies in both languages and hence similar lengths, on the whole. Herdan used information-theoretic measures of correlation thrown up by the theory of discrete communication in the presence of noise. But this was neither essential, nor desirable:such measures are suitable for distributions among qualitative categories (i.e., contingency tables) but are

inadequate for joint distributions of variates, being independent of actual values, absolute or relative.[1]

1.4.28. Herdan (ibid) did not notice the significance of the between language comparisons which are obtained as a bye-product of his work. The average lengths shown below each bivariate distribution lead to estimates of relative efficiencies, ignoring the non-comparability of syllables in different languages. These averages are reproduced below:

language compared with English	passage no.	average length (syllables) per word/ word-group	
		original passage	English translation
(1)	(2)	(3)	(4)
French	1	1.53	1.42
	2	1.47	1.37
German	1	1.74	1.56
	2	1.73	1.58
Czech	1	1.97	1.69
	2	2.05	1.66
Russian	1	2.20	1.64
	2	2.21	1.54

1.4.29. It is desirable that such studies be carried out on larger scale. The present author does not agree with Newman and Waugh (1960) who were misled by their gross underestimates of redundancy and concluded that such comparisons of lengths of original passages and translationsare not very useful or safe.

1.5.1. Indian Work in the Field : A brief account of some Indian work is given here for the sake of interest.

1/ See, in this connection, Linfoot (1957).

1.5.2. Important statistical investigations carried out by Professor S. K. Chatterjee are reported in his treatise on the "Origin and Development of the Bengali Language" (Chatterjee, 1926). Although not employing modern statistical methods, these investigations are convincing, and, as already stated, important.

1.5.3. The vocabulary of Bengali may be classified into some broad etymological groups : (i) 'tatsama' — Sanskrit words either not modified in course of time or late arrivals as loan words, (ii) 'semi-tatsama' — loan words from Sanskrit, partly modified, (iii) 'tadbhava' — Sanskrit words in modified form coming down through all the intermediate languages, (iv) 'desi' — aboriginal words from pre-aryan languages, and (v) 'videsi' — foreign words, especially from Persian and English [op. cit, Part I, App. D]. In the Bengali dictionary by J. M. Das (1916) showing 75000 words and compounds, Chatterjee found 33000 'tatsama' words, 2000 Persian words, 1000 of European origin (including 700 English and 100 Portuguese), the rest being mostly 'tadbhava'. The dictionary probably under-represents the class of 'tadbhava' words. Thus, 'tatsama' and 'tadbhava' words form the great bulk of Bengali vocabulary.

1.5.4. By means of small (non-probabilistic) counts from representative works of different periods, Chatterjee (op. cit, pp.218-23) demonstrated a major phenomenon in the history of the Bengali language. In the 47 "caryapadas" composed in 10-12th centuries, less than 5% out of the 1957 words and compounds fall under the category of 'tatsama',

the remaining 95% were 'tadbhava' (or 'semi-tatsama' or 'desi'); 310 words spell as in Sanskrit but were probably 'tadbhava'. But after this, gradually, 'tadbhava' words were replaced by 'tatsama' forms. Thus, in the 'Srikrishna kirtan' (4th quarter of 14th century), there were 12.5% 'tatsama' words among 863 words selected at random. The proportion of 'tatsama' rose to 25-35% in the 'Rāmayāna' by Krittivasa and other popular works. People became more and more familiar with Sanskrit words under the influence of literature produced by Sanskrit scholars. The extreme was reached in the 19th century by the Fort William group, and by Vidyasagar and Bamkimchandra (early phase). Thereafter the reaction started. In modern works in colloquial Bengali, the percentage of 'tatsama' words is around 20 only. The percentage may be higher, even more than 50, in writings on elevated topics, especially when produced by Sanskrit scholars; but, on the whole, Bengali may be said to have returned to its middle age, in this respect.

1.5.5. Chatterjee (op. cit, pp. 211, 222-3) carried out a similar study on the relative frequency of Bengali words of Persian origin. These were very rare in "Srikrishnakirtan", but became more frequent in course of time, the (relative) frequency being highest in the 18th century, as in "Annadamangal" by Bharatchandra. The frequency declined thereafter. In colloquial Bengali of upper and middle class Hindus in Calcutta, the percentage frequency is generally around 7. In Muslim homes it may rise to even 15, but not 30 as in some artificial "Mussalmani" Bengali.

1.5.6. Finally, Chatterjee (op. cit, pp. 270-4) also studied the relative frequencies of Bengali phonemes. These were estimated from a count of 10,000 phonemes selected from some representative works in prose and poetry. The estimates were compared with those for Sanskrit phonemes given by Whitney (1923) and the divergence between the two distributions was discussed in a very interesting manner.

1.5.7. Deb Chaudhury (1931) carried out a 100,000 word-count in Bengali, mostly in the chaste style, on lines generally similar to those followed by Dewey (1923). The objective was to prepare a Bengali primer on scientific basis. The commonly used primers were shown to be defective by using various statistical criteria, like the size of vocabulary and the frequency of repetitions of words. An improved primer was actually constructed, put to trial and finalised. Unfortunately, only the list of the 1017 commonest words occurring 15 times or more, is shown, without any figure for actual frequencies. (The total number of word-types found was only 6567.) Another list of 509 words is given showing the most frequent words found in a count of 12000 words from juvenile literature; this showed 1627 different words. There is a third list of 1012 words based on the bigger count; this employs a credit system considering range of occurrence in addition to frequency.

1.5.8. Some researches were carried out in the 1940's in the Indian Statistical Institute under the guidance of Professor Mahalanobis. The objectives were threefold : (1) working towards a Basic Bengali based on colloquial Bengali, (2) preparing graded word-lists for use

in text-groups for children in three different age-groups, and (3) studying the vocabularies of representative authors. The most important results were published in Roy (S.N.) and Roy (J.) (1946). The contained, among other things, a list of nearly 3000 words, suitable for the second age-group of children (7-12 years), and detailed instructions for writing text-books for children. During the last few years, again, Roy (J.) has been carrying out extensive word-counts in a unit of the Institute, called the Linguistic Research Unit (LRU). Counts have been completed for the 14 novels by Bankimchandra Chattopadhyay and for the 12 novels and 2 collections of short stories by Rabindranath Tagore. The results are going to be published in the near future.

1.5.9. A statistical study on sentence-length in Kannada prose was made by Subba Rao (1960), who mentions some earlier work by Venkateshiah on the usefulness of lengths of stanzas as indicator of style in poetry. Subba Rao selected eight works by three authors, Kuvempu, Sreenivasa and K. V. Iyer; and for each work he counted lengths of all sentences appearing on about 50 pages selected by systematic sampling. Sentence-length distributions, averages etc., were presented. The average sentence had only about 7 or 8 words in these eight works; the averages found for English prose are usually much higher (vide paras 1.2.16 &1.2.32 above). Sentences with 200 or 300 words are among the longest in English prose as against sentences with 50 words or so in Kannada prose.

1.5.10. Subba Rao compared the authorwise distributions by Kolmogorov-Smirnov test and other techniques. He found significant differences and concluded that sentence-length is an indicator of an author's style. Within author variation was not, however, shown to be negligible, as promised — this has to be done if sentence-length is put forward as a measure of individual style. Secondly, Subba Rao applied methods suitable for unrestricted random sampling without any reserve whatsoever. Finally, certain analyses carried out by him are based on the assumption of lognormality, which should have been verified for his Kannada data.

1.5.11. A recent work by Siromoney (1963) investigates the entropy of Tamil prose. A small-scale experiment on Shannonian lines was carried out : Subjects were asked to guess the next letter when a given number of preceeding letters (upto about 30) were known. The upper bound of the entropy H was estimated as 2.51 bits; here H_o = 4.34 bits, since Tamil has 20 letters, ignoring "Aitham". Relative frequencies of letters were estimated by random sampling methods for modern Tamil prose, and also for poetry of different periods spread over the Christian era. These were used for studying the one-gram entropy H_1.

Chapter 2 : The Samples of Words

2.1.1. Introductory Remarks on Coverage : This chapter describes the material collected for word-length and related studies, the methods of probability and systematic sampling adopted for such purposes, and the properties of estimates obtained from such samples[1].

2.1.2. Initially, sixteen works in Bengali prose, two by each of eight authors, were chosen for the study. These are starred in Table 3.1 of Chapter 3, which covers all the Bengali works included in the word-length studies, excepting some Bengali poems, shown in Table 3.2 of Chapter 3.

2.1.3. About 100 lines were first selected from each of these 16 works strictly at random and with replacement, and all words falling on the selected lines formed the probability samples of words. [For details, see Section 2.3 below.] From a number of these works, then, systematic samples of words were taken in various ways,, e.g., by taking the second line from the bottom of every alternate page or the fourth line from top of every third page. [Vide Section 2.6 below.] It was found that these systematic samples behaved like the probability samples, to a very high degree of approximation; or rather, the systematic samples seemed to have slightly smaller sampling errors. [Vide Section 2.7 below.] That the series of word-lengths is nearly random [vide Chapter 5] was also found at this stage.

---

[1] Chapters 2 to 4 are entirely based on Bengali material. Vide Appendix 3 for corresponding results based on the English novel, "Pride and Prejudice" by Jane Austen.

2.1.4. Many other works were added at a later stage. These were mostly by Bamkimchandra Chatterjee and Rabindranath Tagore, the two greatest writers of Bengali prose. For most of these works, a probability sample was drawn first, but the experimentation with systematic samples was also continued; and since, in general, the systematic samples were found to be valid, the data from the two types of samples were pooled, wherever necessary, for purposes of drawing inferences.

2.1.5. Most of the works covered are novels of different types. The two works by Vidyasagar are free renderings of classical Sanskrit dramas. "Char-Yari Katha" is a string of four short stories; "Chacha Kahini" is a collection of short stories. "Dristipat" and "Deshe Videshe" come under belles lettres. "Birbaler Halkhata" is a collection of essays.

2.1.6. Chronologically, the works cover the entire period of modern Bengali prose. The earliest of them, "Shakuntala" is considered as the first work of art in Bengali prose, and was published in 1854; the latest, "Chacha-Kahini" and "Janantik" came out in 1952. Some of the works represent landmarks in the history of the language and/or literature; certain others, like those by Muztaba Ali and Jajabar, were included as representing distinctive styles or contemporary trends.

2.1.7. Three short essays have also been included, two by Tagore and one by Bamkimchandra. As already stated, the larger

works include "Birbaler Halkhata", a collection of essays by Pramatha Chaudhury. Three short stories by Tagore have also been covered. Two of the larger works, "Chacha Kahini" and "Char-Yari Katha" consist of stories. As regards poetry, the material consisted of the following [vide Table 3.2 of Chapter 3]:

(i) First 200 lines of Canto I of "Meghanadabadha Kavya", the first and most remarkable Bengali epic in blank verse, by Michael Madhusudan Dutta. These lines include both heroic passages and passages full of pathos.

(ii) Twenty-two representative poems of Tagore selected in a purposive manner, so that (a) they are fairly spread over his poetical life-span, (b) represent many of his more famous works and (c) represent many different types of poetry, themes, moods and meters.

2.1.8. The selection of essays, short stories and poems was done very carefully; for while the coverage is small and also the material is largely confined to Tagore, the objective is to generalise for Bengali literature as a whole. It is believed that this generalisation would be more of less valid so far as observations on wordlength and syllable-type are concerned.

2.2.1. The 'Word' and the 'Syllable' : Briefly speaking, words were taken as printed, demarcated from one another by spaces. This is, of course, a limitation of the present study. Bengali is mainly derived from Sanskrit and an appreciable fraction of Bengali words are compounds. Different works use compounds in different proportions.

Compounds are more frequent in the elevated Sanskritised style of Vidyasagar and Bankimchandra (early phase) which has largely gone out of use in modern writings. Between works comparisons of wordlength will not, therefore, have the same meaning as they could have in a language relatively free from compounds[1].

2.2.2. No attempt was made to eliminate this factor by counting compounds of two words, say, as two words, instead of one. The first consideration was that probably the comparisons are more meaningful if no such adjustment is made. One has, of course, to remember that part of the difference between the average wordlengths in Vidyasagar's "Shakuntala" and Tagore's "Ghare Baire" is due to the larger proportion of compounds in the first-named work. The second consideration was the great difficulty involved. Sometimes the component words of a compound are inextricably joined by "Sandhi" (assimilation); in other cases, it is almost impossible to decide whether to treat the word as a compound or not (e.g., for "Pradi" compounds).

2.2.3. Another limitation of partly the same nature must be recorded here. There are a few instances in Bengali, where two consecutive words might be printed either together or with spaces between them, and where the preferences of individual authors vary. Thus, Tagore writes "je-sab", "je-din", where many others would

---

1/ Compounds are also frequent in certain types of poetry, or rather, a high proportion of compounds seems to be quite natural in poetry [vide Chapter 3, Section 3.3].

write "je sab", "je din". These irregularities are also not adjusted for in the present study. The effects are smaller than those of ignoring compounds.

2.2.4. Counting of syllables was based on the standard pronunciation of literary Bengali, which means the modes prevailing in learned circles in and around Calcutta [vide Chatterjee, 1921] In some cases, an 'a' sound ('a' as in English 'fall'), seemed to be optional. Such cases were few in Bengali prose, and the older mode of pronouncing it was adopted there. But the proportion of such cases was larger in Bengali poetry. If the 'a' sound were pronounced, one would get two consecutive syllables of the open or vowel-ending type, while if it were ignored, one got only one closed or consonant-ending syllable[1/]. Consider, for instance, the first word of the opening line of "Meghanadabadha Kavya". It may be pronounced either as "Sammukhasamare", pronouncing the 'a' underlined, or as "Sammukhsamare" without disturbing the meter. In all such cases, the decision had to be taken as to which pronunciation seemed to be more appropriate. For poems in 'payar' and related meters, where the vocal drawl is predominant (Chatterjee, 1945, pp. 379-89) pronouncing the 'a' sound generally seemed to be desirable. It must be admitted that such decisions had to be taken rather too frequently for the word-length and syllable-type data on Bengali poetry to be regarded as absolutely correct.

1/ Vide Chapter 7 for a study on different types of syllables.

2.2.5. The 25 diphthongs of Bengali phonetics [vide Chatterjee, 1945, pp. 34-5] were divided into two groups. The first group comprised

ei, eu, ɔ̃e, ɔ̃eo, ai, ae, ao, au, ɔe, ɔo, oe, oi, ou, ui,

and the second group,

ie, ia, io, iu, ea, eo, ɔa, oa, ue, ua, uo.

The diphthongs in the first group were regarded as similar to single vowel sounds in that they form the core of single syllables; those in the second group were considered as two distinct vowels. All triphthongs and higher combinations were split into different syllables on the basis of the rules adopted for diphthongs[1].

2.2.6. It goes without saying that any English, Sanskrit or Hindi matter found in the Bengali texts, was generally left out, along with poetry passages, if any, found in prose texts. Exceptions were made in a few cases where such matter seemed to be completely integrated with the main text in Bengali prose.

2.3.1. Probability Sampling : The same method of probability sampling was used in all cases, although more convenient methods suggested themselves as the study progressed [see below]. This method consisted in selecting the desired number of lines strictly at random and with replacement, giving at each draw equal probability of selection to each line in the work. All words falling on all the sample

---

1/ The rules described in para 2.2.5 are not completely in accord with linguistic theory, but more or less in keeping with the same, and framed for large scale investigations [vide Chatterjee, 1921, pp.16-17].

lines together formed the probability sample of words[1]. Some hyphenated words occurred partly on the sample lines; so the following convention had to be introduced: exclude the word wholly if it is at the beginning of the sample line and include the word wholly if it is at the end of the sample line.

2.3.2. The method may be regarded as "cluster sampling" of words, lines acting as clusters (Cochran, 1963, Chapter 9; Sukhatme, 1954, Chapter VI). It will be shown in Chapter 5 that the intra-line correlations between lengths of words are quite small, about 0.05 or smaller. So these probability samples of words may be regarded as nearly (unrestricted) random samples of words.

2.3.3. It is, of course, possible to select (unrestricted) random samples by extending the procedure adopted in the present case [see description below]. But the process would be rather too cumbersome and time-consuming, and the simplification of sampling theory not worth the trouble. The "optimum" design seems to be, on retrospect, to use clusters of several consecutive lines as sampling units; the loss of efficiency due to intra-cluster correlations would be more than compensated by the saving in sampling time.

2.3.4. In many cases these probability samples were found to be inconclusive due to small sizes, and systematic samples (vide below) selected from the some works were pooled with the corresponding probability samples. The use of systematic samples seems to be

---

1/ These sample lines were again used for drawing probability samples of sentences [vide Chapter 8].

sufficiently valid for all practical purposes (vide Section 2.7), although in theory, they are always open to some criticism, however weak, because the method followed did not have any element of randomness. What happened is that the systematic samples were generally taken to study their agreement with probability samples — this being an important objective of the present investigation — and when the agreement was found to be really satisfactory, there was little incentive for extending the size of the probability sample by following the original design or otherwise, instead of using systematic samples already selected from the same works.

2.3.5. Probability sampling does not seem to have been used systematically for studies on word-length or sentence-length, or for that matter, for any statistical study on languages[1/]. But it can certainly be used for many such purposes, where it may be unnecessary to exhaust entire works, and where at present "haphazard" or subjectively selected samples are taken [vide Yule, 1938; Elderton, 1949]. While complete counts have their merits [vide Ross, 1950, where the sampling approach is strongly criticised], and "systematic" samples like those used in the present study might give samples which are very nearly random, due to the population sampled being approximately in a random order like the population of word-lengths [vide Chapter 5] or sentence-lengths [Chapter 9], they cannot possibly detract from the usefulness of probability

1/ Vide, however, Chapter 1, paras 1.3.38-41, for discussions on probability sampling carried out by Yule (1938), Subba Rao (1960) and others.

samples. It should also be pointed out that samples comprising "selected" continuous passages are very often used, e.g., by Yule (1938), and such samples, are probably inferior to the systematic samples of the type used here, even though these latter are also non-probabilistic in nature.

2.3.6. The actual procedure of selection seems to deserve mention, inspite of its being obvious to statisticians. The following is the essence of the method; obvious modifications for economic use of random sampling numbers would suggest themselves to statisticians. The selection is done in two stages. First, one page is selected at random. Next, a random number is read between 1 and the maximum number of lines per page in the whole work, or any integer exceeding this. If this random number exceeds the number of lines on the page selected in the first step, the page is rejected, and the process repeated. If the random number is not greater, the corresponding line on the selected page, counting lines from the top of the page downwards, is taken in the sample. For works printed in two columns in each page, a further stage has to be added, namely, the selection of a column with equal probability.

2.3.7. This process is simple and not very time-consuming, but the systematic samples used needed practically no time at all.

2.3.8. The above-mentioned method of selection is related to Lahiri's method (Lahiri, 1951). Suppose Lahiri's method is followed for selecting pages with probability proportional to the number of

lines on the page. When any page happens to be selected by the Lahiri method, the second random number used for selection is here used as the serial number of the sample line within the selected page counting lines from top downwards.

2.4.1. Properties of Probability Sample Estimates : Let n_i be the number of words on the ith randomly selected sample line (i= 1,2,..,k), $n_i^{(r)}$ the number, out of these, of r-syllabled words (r= 1,2,....) and x_{ij} the length in syllables of the jth word on the ith sample line ($j = 1,2,\ldots, n_i$). Then the interest mainly centres upon ratio estimates of the following types :

$$\text{(i)}\quad p_r = \frac{\sum_i n_i^{(r)}}{\sum_i n_i} \text{-(the proportion of r-syllabled words)}$$

$$\text{(ii)}\quad \bar{x} = \frac{\sum_i \sum_j x_{ij}}{\sum_i n_i} \quad \text{(the average word-length)}$$

Here $\sum_i$ denotes the sum over all the k sample lines or clusters.

2.4.2. One can obviously apply the theory of ratio estimates of the form $R = \frac{\sum_i z_i}{\sum_i y_i}$ based on cluster samples (Cochran, 1963, Chapter 6, also Chapter 9, pp. 249-'50; Sukhatme, 1954, Chapter IV, also pp. 265-8). Theoretically, such estimates are consistent, but biased, in general; the bias vanishes if the regression of y on x is a straight line passing through the origin. Series expressions

are available for both bias and sampling variance. The bias is of the order of $\frac{1}{k}$, while the standard error is $O(\frac{1}{\sqrt{k}})$. If k is sufficiently large, the sampling distribution becomes asymptotically normal subject to some mild restrictions on the population sampled; the bias can then be ignored; the first term of the expansion for sampling variance is sufficient; and finally sampling variance can be estimated from the expression

$$V(R) = \frac{1}{k(k-1)\,\bar{y}^2} \sum_{i=1}^{k} (z_i - R\,y_i)^2$$

which is the first term of the expansion just mentioned with sample statistics instead of population values.

2.4.3. Theory also indicates that these large sample results can probably be used if $k \geq 30$ and if further both the sample means $\bar{y}$ and $\bar{z}$ have coefficient of variation below 0.1 (10%).

2.4.4. In the present case the regressions of $n_i^{(r)}$ or $\sum_j x_{ij}$ on n_i resemble straight lines passing through the origin. The biases may therefore be assumed to be negligible. But there is direct evidence to prove this point, and more generally, to show that the large sample results may be used as approximately valid.

2.4.5. First of all, k is at least about 100 for the probability samples from all the works. Secondly, Table 2.1(a) shows the estimated C.V.'s of the sample averages of n_i and of $\sum_j x_{ij}$ obtained from the probability samples of k lines, separately for each of

24 works[1/]. It can be seen that the C.V.'s range from 1.06% to 3.96%. So they are well below 10%, as required for the large sample results to be usable for the estimated averages $\bar{x}$.

Table 2.1(a): Coefficients of variation of sample averages of the number of words (n_i) and the number of syllables ($\sum_i x_{ij}$) on sample lines, estimated from k randomly selected lines comprising probability samples from 24 works in Bengali prose

	name of work	no. of sample lines (k)	estimated C.V.'s (%)	
			C.V.($\bar{n}_i$)	C.V.($\bar{\sum_j x_{ij}}$)
	(1)	(2)	(3)	(4)
1.	Shakuntala	100	2.06	1.89
2.	Sitar Vanavas	100	1.44	1.06
3.	Durgeshnandini	100	2.53	2.29
4.	Visavriksha	99	2.80	2.87
5.	Krishnakanter Will	200	2.52	2.64
6.	Anandamath	200	2.70	2.57
7.	Devi Choudhurani	200	2.13	2.12
8.	Rajsinha	250	1.74	1.78
9.	Bouthakuranir Hat	200	2.15	2.12
10.	Rajarsi	200	1.92	1.92
11.	Gora	100	2.80	2.75
12.	Chaturanga	200	1.95	1.92
13.	Ghare Baire	200	1.93	1.87
14.	Sheser Kavita	100	2.92	2.77
15.	Char-Yari Katha	100	3.12	3.23
16.	Birbaler Halkhata	100	2.40	2.04
17.	Pallisamaj	100	2.74	2.70
18.	Pather Dabi	100	3.30	3.46
19.	Pather Panchali	100	2.61	2.67
20.	Devayan	100	3.29	3.45
21.	Dristipat	100	2.42	2.32
22.	Janantik	100	3.96	3.26
23.	Deshe Videshe	100	2.89	3.08
24.	Chacha Kahini	100	3.11	3.19

1/ The correlation coefficient between n_i and $\sum_j x_{ij}$ was found to be nearly 0.85, on an average, and was higher for works with higher values of C.V.(n_i) or C.V.($\sum_j x_{ij}$), which indicates the relative frequency of incomplete or broken lines. Also, the two latter-mentioned C.V.'s were nearly equal, in general.

Table 2.1(b): Coefficients of variation and other characteristics of the $n_i^{(r)}$ for r = 1,2,..., 7 estimated from 200 randomly selected lines comprising the probability samples from each of "Anandamath"and"Ghare Baire"

	word-length in syllables (r)						
	1	2	3	4	5	6	7
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C.V.'s (%) of $n_i^{(r)}$							
1. Anandamath	109.99	68.12	66.96	119.3	337.6	529.1	700.0
2. Ghare Baire	74.86	40.38	79.20	150.4	307.1	700.0	1410.6
C.V.'s (%) of sample averages of $n_i^{(r)}$							
3. Anandamath	7.78	4.82	4.73	8.44	23.87	37.41	49.50
4. Ghare Baire	5.29	2.86	5.60	10.64	21.72	49.50	99.74
means of $n_i^{(r)}$							
5. Anandamath	0.900	2.255	1.665	0.560	0.110	0.035	0.020
6. Ghare Baire	1.965	5.440	1.515	0.445	0.105	0.020	0.005
estimates of p_r							
7. Anandamath	16.23	40.67	30.03	10.10	1.98	0.63	0.36
8. Ghare Baire	20.67	57.23	15.94	4.79	1.10	0.21	0.05

2.4.6. As regards the estimates p_r, the random variable in the denominator (n_i) has just been disposed of, and Table 2.1(b) shows the C.V.'s of the sample averages of the numerator variable $n_i^{(r)}$, for r = 1,2,...,7, estimated from the probability samples from two selected works. The condition that the C.V. of the sample mean should be less than 10 per cent is fulfilled for $n_i^{(1)}$, $n_i^{(2)}$ and $n_i^{(3)}$, and also, nearly, for $n_i^{(4)}$, but not for $n_i^{(r)}$ with $r \geq 4$. Hence, the large sample properties may be assumed for the estimates p_1, p_2, p_3

and (probably) p_4, but not for p_5, p_6 etc. Since "Ghare Baire" has a low average of word-length and "Anandamath" a rather high average [vide Table 3.1 of Chapter 3], Table 2.1(b) should indicate the general picture. But there is one important point. Whereas the C.V.'s in rows 1 and 2 are generally meaningful, those in rows 3 and 4 are not, having been obtained by dividing corresponding figures in rows 1 and 2 by $\sqrt{200}$, 200 being the number of sample lines from either work. For many works, e.g.,"Pallisamaj", only 100 sample lines were selected for the probability sample; in such cases, the divisor would be only $\sqrt{100}$ and the large sample assumption would be unsafe even for p_4. It may be stated, however, that for most of the works on which the main findings are based, 200 or more sample lines were used, considering probability and systematic samples of words together.

2.4.7. The sample standard deviation of x uses the ratio $\frac{\sum_i \sum_j x_{ij}^2}{\sum_i n_i}$ in addition to the ratio $\bar{x}$ already discussed. This ratio would probably have properties similar to those of $\bar{x}$, and so the estimates of s.d. may be taken to be practically unbiased etc.

2.4.8. Certain other types of estimates will be introduced in Chapters 4, 5 and 7. The relevant discussion will be found in the appropriate chapters.

2.5.1. Interpenetrating Subsamples (IPNS) : The sample of k randomly selected lines was split up into four[1/] independent and interpenetrating subsamples (ss) : SS 1 comprised sample lines numbered 1, 2, ..., $\frac{k}{4}$, in the order of selection; SS 2 lines $\frac{k}{4}$ + 1, ..., $\frac{k}{2}$; and so on. The estimates (p_r, $\bar{x}$ etc.) were obtained separately by subsamples and also for the full or combined sample. The subsample estimates have the same ratio form but are based on $\frac{k}{4}$ pairs (y_i, z_i) only.

2.5.2. The question arises as to how far these subsample estimates are unbiased and possess the other large sample properties. The sample size $\frac{k}{4}$ may be as low as 25. But so far as the estimates $\bar{x}$ are concerned the position is probably satisfactory, since the two C.V.'s in Table 2.1(a) would remain less than 10% even when multiplied by 2, so that they relate to the four subsamples. From Table 2.1(b), however, it appears that the subsample estimates p_r may not possess the large sample properties for even the low values of r, unless k is at least 200. Using two halfsamples instead of four subsamples would be convenient in this respect.

2.5.3. That even the subsample estimates based on $\frac{k}{4}$ observations — $\frac{k}{4}$ may be as low as 25 — are more or less unbiased can be seen indirectly from Table 2.2. This table compares the "combined"

1/ For a number of works 8 or 10 subsamples were used for certain special purposes [vide Section 2.7, Table 2.4(a)], while for other purposes the subsamples were lumped together to give two halfsamples (vide Section 2.7, Table 2.3 , Fig. 2.1)

estimates $\bar{x}$ and p_r (r = 1,2,...., 6), based on all the subsamples taken together, with the corresponding simple averages of the four subsample averages. [The subsamplewise estimates may be seen in Tables 3.1 and 3.3 in Chapter 3.] The combined estimate, which is a weighted average of the subsample estimates, agrees very closely with the simple average of the subsample estimates. Table 2.2 shows the differences in percentage terms. In anticipation of conclusions to be reached in Section 2.7, the table also includes the comparisons for the systematic sample estimates.

2.5.4. Col.(11) of Table 2.2 shows that, so far as the estimates $\bar{x}$ are concerned, the percentage difference exceeds 0.1 in only 14 rows out of 48; and the maximum difference is only -0.726 per cent. The differences are thus negligible for practical purposes. [There is, however, a significant tendency of the percentages being positive.] The differences are also negligible for the estimates p_r, though not always in the absolute sense, at least for p_6 [col.(10)], where the estimates themselves are small in numerical value.

Table 7.1: Differences between the simple averages of subsample estimates and the 'combined' (all subsample) estimates expressed as percentages of the latter.

name of work	type of sample	no. of subsamples	total no. of sample words	differences (%) for proportions of words of different lengths in syllables. 1	2	3	4	5	6	for average word-length in syllables
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Shakuntala	prob.	4	696	-1.14	-0.29	-0.31	0.64	-0.22	2.17	0.218
2. Sitar Vanavasa	prob.	4	750	-0.22	-0.07	0	0.04	0.50	0.62	0.070
3. Durgeshnandini	prob.	4	577	-1.17	-0.37	0.43	0.57	1.38	-0.83	0.345
	syst.	4	1782	-0.44	0.11	0.44	0.39	-0.28	0.95	0.002
4. Kapalkundala	syst.	4	493	-2.29	0.84	-0.43	1.10	0.60	2.09	0.250
5. Visavriksha	prob.	4	611	-0.52	0.25	-0.17	-0.47	1.76	1.92	0.093
	syst.	4	1852	-0.11	-0.07	0.10	0.13	-0.07	-0.66	0.029
6. Krishnakanter Will	prob.	8	1777	-0.28	-0.04	0.19	0.23	-0.27	0	-0.039
	syst.	4	749	0.19	-0.04	-0.34	0.37	0.07	5.62	0.046
7. Anandamath	prob.	8	1109	0.01	-0.04	0.13	-0.20	0.06	-0.40	0.030
	syst.	4	801	-0.67	0.11	-0.20	1.02	1.12	-1.61	-0.726
8. Devi Choudhurani	prob.	8	1174	-0.40	-0.19	-0.61	0.86	1.47	4.81	0.095
	syst.	4	833	-0.98	0.03	0.26	1.31	2.83	-	0.250
9. Rajsinha	prob.	10	1423	0	0.03	-0.88	0.21	0.34	0	0.022
	syst.	4	507	-0.21	0.14	-0.12	0.23	-0.80	0.64	-0.013
10. Bouthakuranir Hat	prob.	8	1592	0.19	-0.36	-0.50	-0.20	0.77	0	0.030
	syst.	4	827	-0.55	-0.26	0.44	0.24	0.28	3.33	0.145
11. Rajarsi	prob.	8	1632	0.25	-0.16	0.04	0.35	-0.61	0.45	-0.016
	syst.	4	689	-0.53	0.39	-0.52	1.00	0	-4.11	0.013
12. Chokher Bali	syst.	4	1318	-0.36	-0.03	-0.01	0.76	0	3.33	0.101

Table 2.2: (Contd.)

name of work	type of sample	no. of sub-samples	total no. of sample words	differences (%) for proportions of words of different lengths in syllables						for average word-length in syllables
				1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
13. Gora	prob.	4	889	0.03	-0.22	0.17	0.64	0.53	-4.55	0.060
	syst.	4	1824	0.39	0.10	-0.55	0.75	-0.42	0.93	-0.058
14. Chaturanga	prob.	8	1458	-0.39	-0.10	0.15	1.64	-1.15	-4.47	0.109
	syst.	4	854	0.17	-0.17	0.23	-0.19	-0.61	1.09	-0.010
15. Chare Baire	prob.	8	1901	-0.68	0.04	0.28	1.64	-0.34	2.38	0.104
16. Shesher Kabita	prob.	4	735	-0.33	0.06	0.01	0.28	0.86	0	0.073
	syst.	4	1284	**-0.27**	**0.03**	**0.05**	**0.78**	**-0.55**	**2.52**	-0.201
17. Yogayog	syst.	4	1187	-0.06	-0.01	0.08	-0.08	0	1.19	0.014
18. Char-Yari Katha	prob.	4	872	0.09	-0.08	0.14	0.05	0.31	-1.63	-0.005
19. Birbaler Halkhata	syst.	4	1041	-0.10	0.01	0.05	0.14	-1.30	3.36	0.043
20. Jalli Samaj	prob.	4	890	-0.14	0.03	-0.03	-0.38	0.35	4.55	0.041
21. Pather Dabi	prob.	4	815	0.11	0.25	-0.46	-0.14	-1.02	6.25	-0.072
22. Pather Panchali	prob.	4	922	0.05	-0.10	1.37	-0.40	-0.20	-	-0.009
	syst.	4	1630	**0.01**	**0.05**	**0.18**	**-0.29**	**0.67**	**3.27**	0.090
23. Aparajita	syst.	4	1894	0.06	-0.15	0.07	0.13	3.24	-1.56	0.097
24. Devayan	prob.	4	931	-0.06	-0.02	0.12	-0.12	1.07	-4.55	0.024
	syst.	4	2245	-0.01	0.02	0.07	-0.21	0	-1.39	-0.004
25. Dristipat	prob.	4	772	0	-0.07	0.22	-0.22	0.17	-	0.008
	syst.	4	1591	0.06	0.09	-0.15	-0.16	-0.09	-0.50	0.292

Table 2.2 (Contd.)

name of work	type of sample	no. of sub-samples	total no. of sample words	differences (%) for proportions of words of different lengths in syllables						for average word-length in syllables
				1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
26. Janantik	prob.	4	690	-0.49	-0.20	0.49	0.49	1.23	2.05	0.174
27. Chacha Kahini	prob.	4	778	0.27	-0.03	-0.16	-0.34	1.36	1.92	-0.009
28. Deshe Videshe	prob.	4	791	-0.62	0.02	0.48	0.73	-1.15	-	0.101
29. Samya	syst.	4	1010	-1.56	-0.12	0.67	0.34	0.44	4.70	0.305
30. Bamkimchandra	syst.	4	1237	0.22	0.04	-0.16	-0.15	-0.26	2.33	0.009
31. Vishwavidyalay	syst.	4	1009	-0.22	-0.10	0.18	0.36	0.40	0.95	0.085
32. Kabuliwalla	syst.	4	779	-0.15	0.01	0	0.05	0.11	0.78	0
33. Kshudhita Pasan	syst.	4	1192	-0.27	0.13	-0.15	0.17	0.45	0.82	0.030
34. Laboratory	syst.	4	1228	0.52	-0.21	0	0.29	1.27	-4.55	-0.035

2.5.5. The idea behind Table 2.2 is this: the bias of the ratio estimate $r = \frac{\sum z_i}{\sum y_i}$ based on k pairs of observations is known to be of the order of $\frac{1}{k}$ ignoring terms of higher order. One may then write

$$E(r) = A + \frac{B}{k}$$

ignoring terms of higher order. Here E denotes expectation, A the true value of the ratio and B a constant depending on the moments of the joint distribution of y and z. The bias of the subsample estimates r_1, r_2, r_3 and r_4 should therefore be $\frac{4B}{k}$, ignoring terms of higher order, and the same must be true of their simple average denoted by r'. To order $\frac{1}{k}$, then

$$E(r') = A + \frac{4B}{k} .$$

It follows that to the same degree of approximation

$$E(r' - r) = \frac{3B}{k} = 3 \text{ (bias of } r \text{ itself)}.$$

Hence, one third of the difference $r' - r$ estimates, to order $\frac{1}{k}$, the bias of the latter estimate, apart from fluctuations due to sampling, and $\frac{4}{3}$ times this difference estimates the bias of the subsample estimates r_i ($i = 1, 2, 3, 4$)[1/].

2.5.6. The agreement between the combined (all-subsample) estimates and the simple averages of subsamplewise estimates shown in Table 2.2, therefore implies that the subsample

---

1/ This fact was exploited by Murthy and Nanjamma (1959) for obtaining 'almost unbiased' ratio estimates.

estimates as well as the 'combined' ones are all practically unbiased. For the estimates of $\bar{x}$ at least, bias is generally less than 10% of the standard error, so that the effect of bias may be taken as negligible.

2.5.7. Assuming that the four subsamples give unbiased estimates of the population ratio, one might estimate the standard error of the 'combined' estimate r as $\frac{1}{\sqrt{4}}\sqrt{\frac{\sum(r_i - r)^2}{3}}$. If normality is assumed, one can set up confidence limits using Student's t with 3 degrees of freedom. [A more general method is indicated in Cochran, 1963, pp. 164- 5.]. Alternatively, assuming merely that the subsample estimates are symmetrically distributed around the true value, one might use the confidence limits $[\min_i r_i, \max_i r_i]$, which include the true value with probability $1-(\frac{1}{2})^3 = \frac{7}{8}$ [Lahiri, 1954, 1957].

2.5.8. The calculation of standard error using the formula given in para 2.4.2 was done only for the estimated averages $\bar{x}$; for the proportions p_1, p_2, p_3 .... of words having lengths 1, 2, 3, ... this type of calculation was not done, and the subsample estimates were used for drawing statistical inferences of a broad nature.

2.6.1. Systematic Samples : The systematic samples were generally drawn in the following manner. The 4th line, from the top of the page was selected from every odd-numbered page, or the 3rd line from

the bottom was chosen from every fifth page[1]. Quite often the sample lines were selected by using more than one such numerical rule for sampling from a given work. Words falling on selected lines comprised the "systematic" samples of words.

2.6.2. The term "systematic" seems to be quite appropriate here, even though the intervals between successive sample lines do, in general, vary, since all pages of a given work do not have the same number of lines. Such small deviations are allowed in other cases also[2]. In the present case, such deviations can only improve the chances of the sample being representative.

2.6.3. For the sampling fractions frequently used in the present study, systematic samples with a fixed length of interval between successive sample lines would have taken considerable time for their selection. For instance, if one had to select every 40th line, the counting of lines would have been quite a problem. The sampling fractions were high only for the three short essays and the three short stories shown at the end in Table 3.1; and in these cases, the interval-length was kept strictly constant, by taking every 3rd or 10th line (say) in the sample.

1/ The first and the last lines of the pages were usually avoided in view of the possibility that where chapters begin on fresh pages, such lines may conceivably have certain peculiarities. But if this were true, avoiding them altogether cannot also be satisfactory although the biases should be smaller in that case.

2/ See, for example, Cochran, 1963, p. 206, regarding card sampling from a drawer.

2.6.4. No use was made of any kind of random start, even where the interval-length was kept strictly equal. Indeed, the samples do not involve any element of probability.

2.6.5. The lines constituting the systematic sample from any work were divided into four interpenetrating subsamples. Suppose the sample lines are numbered 1, 2, 3, ..., according to their position in the natural reading order. Subsample 1 comprised sample lines numbered 1, 5, 9, ..., subsample 2, lines 2, 6, 10, ..., and so on. Estimates were prepared separately for the subsamples as well as for the 'combined' (i.e., full) sample.

2.6.6. Strictly speaking, one cannot think of sampling errors of estimates based on systematic samples where no element of randomness is involved; and even if a random start were made, the estimation of sampling errors would be extremely difficult (theoretically impossible) [Cochran, pp. 224- 7; Sukhatme, pp.431- 3] Our approach to the problem is based on the finding that the series of word-lengths is very nearly random [vide Chapter 5]. The literature on systematic sampling[1] encouraged us to take a very "practical" view of the situation, and to decide to assess the sampling errors of systematic sample estimates by the differences among the subsample estimates. One may imagine that the whole work is

---

1/ [Vide Cochran, 1963, Section 8.10; Sukhatme, 1954, pp.431-3] Note, for example, the discussion of cases where the population is in random order; see also para entitled "Stratification effects only" on p. 225 of Cochran (1963).

vided into a number of strata, and the four subsamples each select one line from every stratum. The same type of idea appears in Mahalanobis [1958, para 7.3] where fractile graphs based on time series are defined.

2.7.1. Validity of systematic Samples : The general agreement between probability samples and systematic samples drawn from the same work will be evident from the word-length data presented in Chapters 3 and 4. For the sake of interest, the length distributions of sample words are shown in Table 2.3 for four selected works, separately for probability and systematic samples. The distributions are expressed by decile group averages[1], and the four [or more] subsamples of either type of sample are pooled to give two halfsamples in every case. Fig. 2.1 shows these averages in four sets of fractile graphs, there being one set for each work. These should serve to show the agreement between the two types of samples.

2.7.2. The examination of fractile graphs was supplemented by some tests of significance. This was because we wanted to establish the validity of systematic samples in a more objective and conclusive manner. There are cases [cf. the estimates for "Rajsinha" in Table 3.1] where the systematic sample estimates deviate considerably from those based on probability samples. But one or two such deviations should rather be expected to appear by chance when a large number of

---

1/ Vide Mahalanobis (1958, 1960) for description and uses of fractile estimates and graphs.

comparisons are made between the two types of samples. It is therefore necessary to carry out an 'overall' test for deciding whether the frequencies of large and small deviations between the two types of samples are more or less as could be expected to occur by chance. Second, the interest is not merely in seeing whether nearly equal estimates of $\bar{x}$ or p_r are being thrown up by the two types of samples. It is also necessary to get sure whether the sampling errors of the two sets of estimates are also more or less equal, apart from differences in the respective sample sizes.

2.7.3. Tables 2.4(a) and 2.4.(b) summarise four series of tests.

2.7.4. Consider, first, the test covered in cols. (2) - (6) of Table 2.4(a). For each work from which a probability sample was taken, the frequency distribution of sample words by length in syllables was prepared separately for the 4 or more subsamples. This gave a two-way contingency table, rows representing subsamples and columns, word-length in syllables. [The material is presented in Table 3.3 in the percentage form.] The χ^2 test of homogeneity was then applied, after pooling some of the higher length classes, if necessary, to avoid small expected frequencies. In spite of Cochran's recommendations (Cochran, 1952, 1954)[1] the minimum expectation was taken as 4; but if 4-syllabled and 5-syllabled words were going to be pooled by following this rule, it was permitted to go down to expected frequencies upto or a little below 3. Similar tests of homogeneity were applied to compare the four subsamples of the systematic sample. The same procedure was followed regarding pooling of small frequencies. The results are presented in cols.(5) and (6) of Table 2.4(a).

1/ As these tests are very important for the rest of the investigation, no risk was taken by using very small expected frequencies.

Table 2.3: Decile grpup averages of word-length in syllables, separately by halfsamples and combined, based on probability and systematic samples of words from four selected works in Bengali prose.

name of work	type of sample	half-sample	no. of sample words	decile group (per cent) 0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	0-100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1. Anandamath	prob.	1	403	1.000	1.586	2.000	2.000	2.000	2.343	3.000	3.000	3.390	4.871	2.519
		2	398	1.000	1.040	2.000	2.000	2.000	2.000	2.819	3.000	3.256	4.552	2.367
		comb.	801	1.000	1.315	2.000	2.000	2.000	2.082	3.000	3.000	3.323	4.663	2.438
	syst.	1	580	1.000	1.397	2.000	2.000	2.000	2.155	3.000	3.000	3.258	4.378	2.419
		2	529	1.000	1.355	2.000	2.000	2.000	2.480	3.000	3.000	3.362	4.493	2.469
		comb.	1109	1.000	1.377	2.000	2.000	2.000	2.310	3.000	3.000	3.307	4.432	2.443
2. Gora	prob.	1	434	1.000	1.598	2.000	2.000	2.000	2.000	2.640	3.000	3.000	3.990	2.323
		2	455	1.000	1.308	2.000	2.000	2.000	2.000	2.736	3.000	3.033	4.308	2.339
		comb.	889	1.000	1.448	2.000	2.000	2.000	2.000	2.690	3.000	3.000	4.172	2.331
	syst.	1	931	1.000	1.432	2.000	2.000	2.000	2.000	2.888	3.000	3.020	4.322	2.366
		2	893	1.000	1.399	2.000	2.000	2.000	2.000	2.449	3.000	3.053	4.338	2.324
		comb.	1824	1.000	1.416	2.000	2.000	2.000	2.000	2.673	3.000	3.036	4.329	2.345
3. Sheser Kavita	prob.	1	366	1.000	1.279	2.000	2.000	2.000	2.000	2.000	2.760	3.000	4.204	2.224
		2	369	1.000	1.000	1.913	2.000	2.000	2.000	2.000	2.574	3.000	4.002	2.149
		comb.	735	1.000	1.095	2.000	2.000	2.000	2.000	2.000	2.666	3.000	4.098	2.186
	syst.	1	647	1.000	1.223	2.000	2.000	2.000	2.000	2.000	2.859	3.000	3.972	2.205
		2	637	1.000	1.132	2.000	2.000	2.000	2.000	2.000	2.701	3.000	4.378	2.221
		comb.	1284	1.000	1.178	2.000	2.000	2.000	2.000	2.000	2.781	3.000	4.086	2.205
4. Pather Panchali	prob.	1	461	1.000	1.308	2.000	2.000	2.000	2.000	2.514	3.000	3.000	4.150	2.297
		2	461	1.000	1.000	1.983	2.000	2.000	2.000	2.000	2.950	3.000	4.106	2.104
		comb.	922	1.000	1.143	2.000	2.000	2.000	2.000	2.232	3.000	3.000	4.127	2.250
	syst.	1	858	1.000	1.159	2.000	2.000	2.000	2.000	2.276	3.000	3.000	4.075	2.251
		2	772	1.000	1.471	2.000	2.000	2.000	2.000	2.549	3.000	3.000	4.080	2.310
		comb.	1630	1.000	1.307	2.000	2.000	2.000	2.000	2.405	3.000	3.000	4.080	2.279

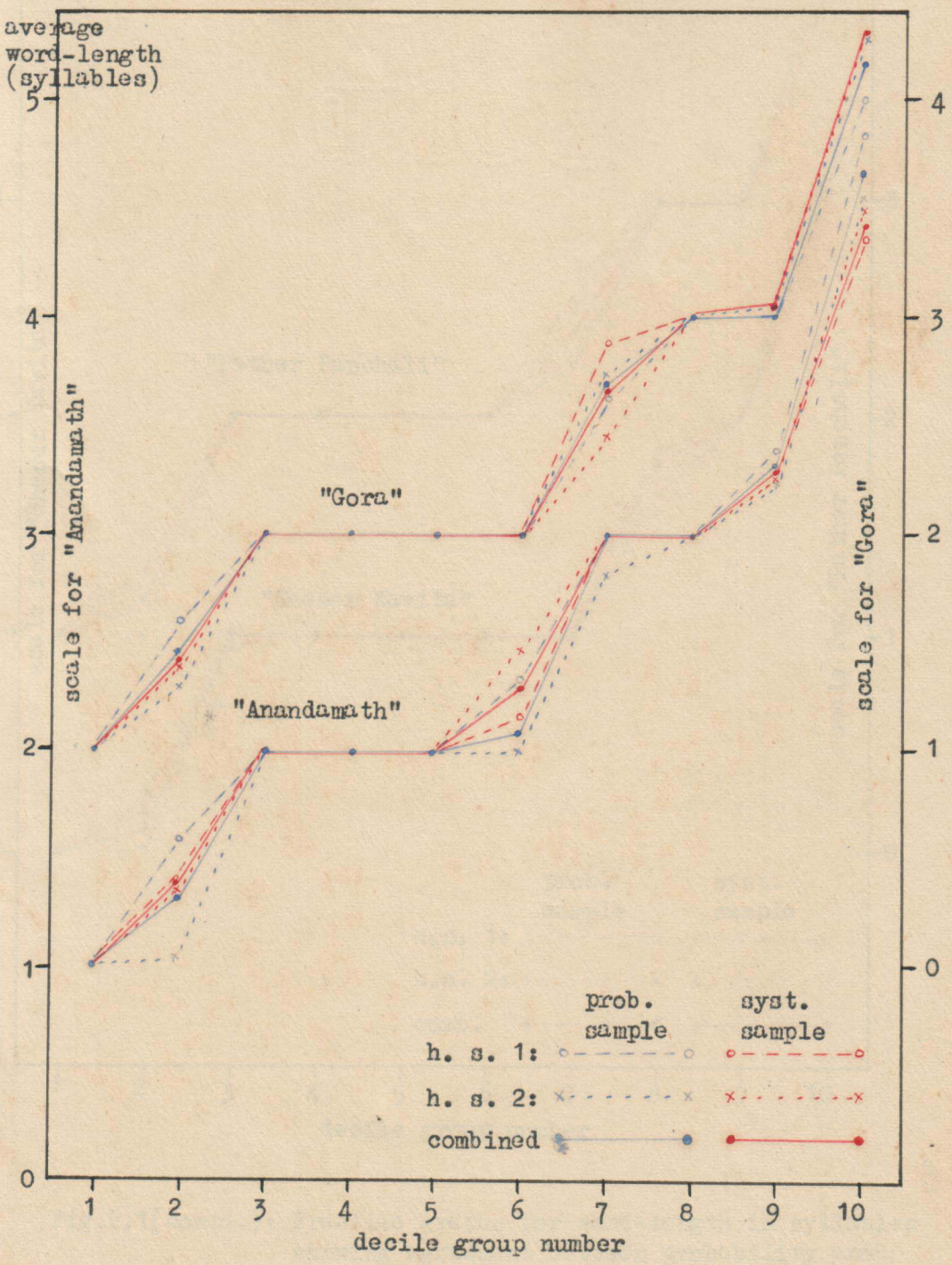


Fig.2.1: Fractile graphs for word-length in syllables showing agreement between probability and systematic samples from four selected works in Bengali prose.

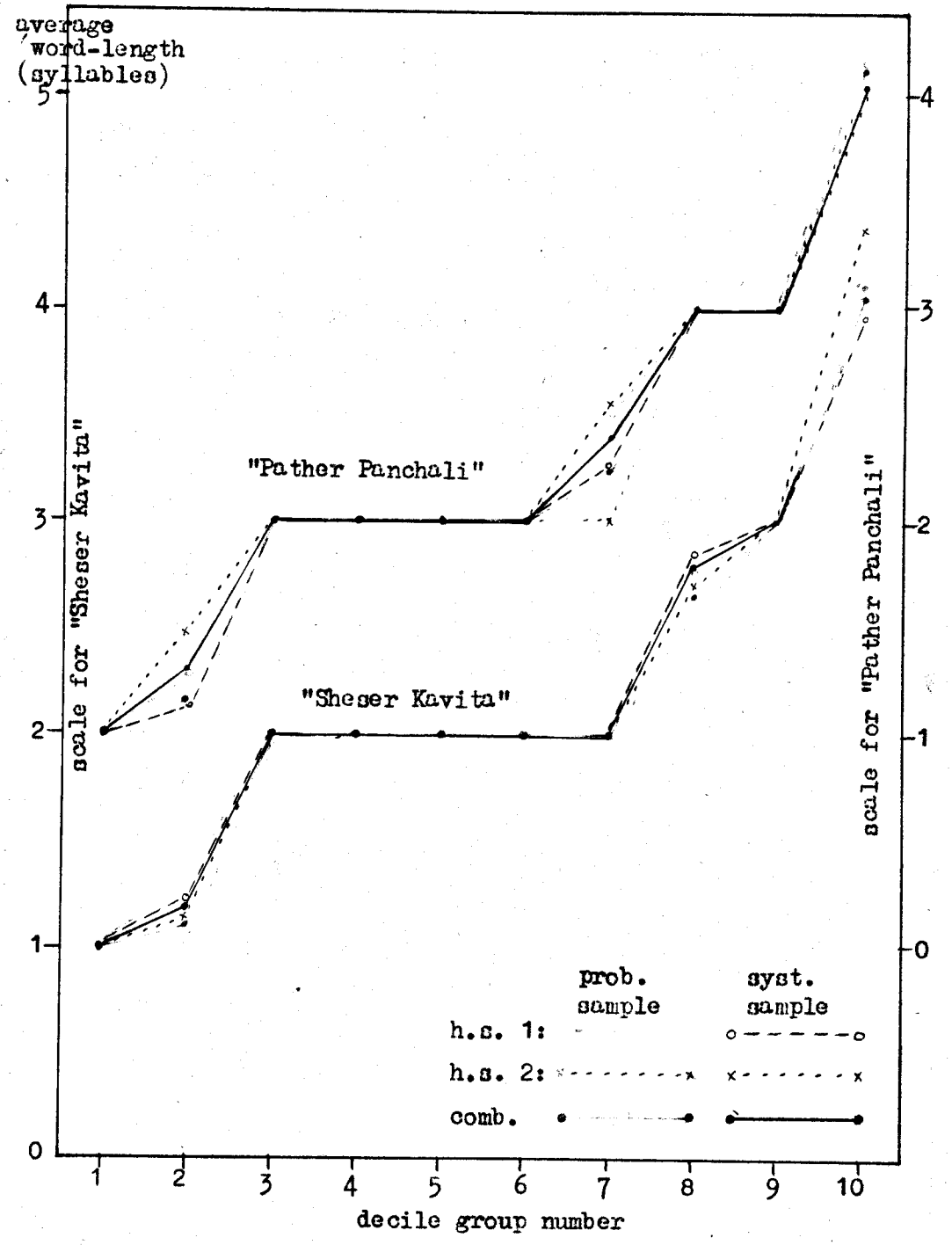


Fig.2.1(contd.): Fractile graphs for word-length in syllables showing agreement between probability and systematic samples from four selected works in Bengali prose.

Table 2.4(a). Results of χ^2 tests for homogeneity of subsample distributions of word-length

name of work	χ^2 for testing homogeneity of word-length distributions based on subsamples of probability samples			four subsamples of systematic samples		combined probability and combined systematic samples	
	no.of sub-samples	d.f.	χ^2	d.f.	χ^2	d.f.	χ^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Shakuntala	4	12	18.509				
2. Sitar Vanavas	4	15	14.140				
3. Durgeshnandini	4	12	17.135	15	15.964	5	3.291
4. Kapalkundala				12	14.809		
5. Visavriksha	4	12	12.623	15	12.611	5	1.792
6. Krishnakanter Will	8	28	39.779	12	8.354	5	5.072
7. Anandamath	8	28	21.874	12	14.015	5	4.832
8. Devi Choudhurani	8	21	29.088	12	19.126	4	2.165
9. Rajasinha	10	36	35.260	12	11.139	5	5.867
10. Bouthakuranir Hat	8	28	20.443	12	16.196	4	3.573
11. Rajarsi	8	28	23.226	12	20.995	5	3.339
12. Chokher Bali				12	11.713		
13. Gora	4	12	16.335	12	21.724*	4	1.440
14. Chaturanga	8	28	41.812*	12	13.856	4	1.347
15. Ghare Baire	8	21	26.194				
16. Sheser Kabita	4	12	18.621	12	11.052	5	5.674
17. Yogayog				12	7.957		
18. Char-Yari Katha	4	12	13.577				
19. Birbaler Halkhata	4	12	7.282				
20. Pallisamaj	4	12	10.153				
21. Pather Dabi	4	9	14.462				

Table 2.4(a) : (Contd.)

name of work	χ^2 for testing homogeneity of word-length distributions based on subsamples of probability samples			four subsamples of systematic samples		combined probability and combined systematic samples	
	no. of subsamples	d.f.	χ^2	d.f.	χ^2	d.f.	χ^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
22. Pather Panchali	4	12	13.138	12	18.900	4	2.917
23. Aparajita				12	13.800		
24. Devayan	4	12	14.703	12	12.473	4	1.436
25. Dristipat	4	12	10.581	15	15.122	5	4.267
26. Janantik	4	12	16.808				
27. Chacha Kahini	4	9	7.152				
28. Deshe Videshe	4	12	5.589				
29. sub-total (1-28)	-	407	448.484	225	259.808	64	47.012
30. Samya				15	20.140		
31. Bamkimchandra				15	34.091**		
32. Vishwavidyalay				12	18.869		
33. Kabulivalla				12	5.200		
34. Kshudhita Pasan				15	12.750		
35. Laboratory				12	17.152		
36. sub-total (30-35)				81	108.202		
37. total (29 + 36)	-	407	448.484	306	368.010	64	47.012

N.B.: (1) Asterisk denotes significance at 5% level and double asterisk at 1% level.
(2) Systematic sampling was slightly different in the two sub-sets of works; also the sampling fraction was appreciably higher in works numbered 30-35 [vide Section 2.6]

Table 2.4 b) χ^2 test for comparing the variability of four subsample averages from the systematic sample with that of the combined average from the probability sample after adjusting for differences in sample size, separately for 14 works in Bengali prose.

name of work	no. of sample words		χ^2 value (3 d.f.)	upper tail probability (P)
	prob. sample	syst. sample		
(1)	(2)	(3)	(4)	(5)
1. Durgeshnandini	577	1782	0.009	>0.99
2. Visavriksha	611	1852	1.793	0.50 - 0.70
3. Krishnakanter Will	1777	749	0.698	0.80 - 0.90
4. Anandamath	1109	801	8.294	0.02 - 0.05
5. Devi Choudhuran8	1174	833	5.943	0.10 - 0.20
6. Rajsinha	1423	507	3.275	0.30 - 0.50
7. Bouthakuranir Hat	1592	827	6.193	0.10 - 0.20
8. Rajarsi	1632	689	4.525	0.20 - 0.30
9. Gora	889	1824	1.977	0.50 - 0.70
10. Chaturanga	1458	854	0.903	0.80 - 0.90
11. Sheser Kabita	735	1284	2.724	0.30 - 0.50
12. Pather Panchali	922	1630	2.776	0.30 - 0.50
13. Devayan	931	2245	1.491	0.50 - 0.70
14. Dristipat	772	1591	5.712	0.10 - 0.20
15. total	-	-	46.313 (42 d.f.)	0.20 - 0.30

* See para 2.7.8 of text, for explanation.

2.7.5. Only a small fraction of these homogeneity χ^2's are significant, 1 out of 24 in col. (4) and 2 out of 24 in col. (6). Apparently, the P-values for either series (not shown) seem to be fairly spread over the interval (0.1). Thus, in col. (6), while two values are significant at 5% level, one falls below the lower 5% level; in col. (4) also, one χ^2 is close to the lower 5% point. But combination of tests gives a different overall picture. If one combines the χ^2's by straight addition[1], one gets, for col. (4), a total of 448.484 which is a χ^2 with 407 d.f. This is not quite significant at the 5% level, but seems to be rather high (P = 0.08 to be precise, using the Wilson-Hilferty approximation). The corresponding total for col.(6) is 368.010 and the degrees of freedom total upto 306. This gives P=0.009, so that χ^2 is here highly significant. So it might be concluded that the χ^2's in col.(4) and (6) seem to have some upward bias, which is not significant for most of the individual works but which appears — significantly for col. (6) — when either series of χ^2's is considered as a whole.

2.7.6. One may now turn to cols. (7) and (8) of Table 2.4(a) relating to the third series of χ^2 tests. There were 14 works in all where both types of sampling had been used. For each of these works, the two word-length distributions, one based on the 'combined' probability sample and the other based on the 'combined' systematic

---

1/ See Bhattacharya (1961) where it was found that for combining a number of independent χ^2 tests, the technique of addition of χ^2's and the P_λ-technique are almost equally powerful.

sample, were brought together in the form of a 2 x p contingency tables. The χ^2 test of homogeneity was then applied, after pooling of small frequencies, if necessary. The minimum expection was taken as 5, but if length classes 5 and 6 were going to be pooled, expected frequencies as low as 4 were permitted.

2.7.7. It can be seen that in not a single case, the χ^2 value in col.(8) reaches even the 30% point. On the other hand, there are values with the P-value near 0.90. Within this range, the P-values seem to be fairly well spread. The straight sum of the 14 χ^2's is 47.012. This is a χ^2 with 64 degrees of freedom. The P-value is 0.953. This means significant evidence that these fourteen χ^2's are a little on the low side.

2.7.8. Another series of tests may be described before interpreting the results. These tests were also applied for each of the 14 works where both methods of sampling were used. The results are shown in Table 2.4(b). The objective was to see whether the variability of the four subsample averages $\bar{x}_1$, $\bar{x}'_2$, $\bar{x}'_3$ and $\bar{x}'_4$ based on the systematic sample is equal to the variability of the corresponding averages from the probability samples, except for differences in sample size. Sample size was measured by number of words; number of lines would make little difference. One may assume $\bar{x}'_1$, ..., $\bar{x}'_4$ are independently and normally distributed, and that the standard error of such estimates for either type of sample is inversely

proportional to the square root of sample size with the constant of proportionality same for both types of sample. One will then see that

$$\chi^2 = \frac{n'}{4n} \quad \frac{\sum_{i=1}^{4} (\bar{x}_i' - \bar{x}')^2}{\text{Est. Var. of } \bar{x}}$$

would be approximately distributed as a χ^2 with 3 degrees of freedom. Here n, n' are the sizes of the (combined) probability and the (combined) systematic samples, $\bar{x}$, $\bar{x}'$ the corresponding combined means, and variance of $\bar{x}$ is estimated by using the expression given in para 2.4.2 so that it can be taken as nearly exact.

2.7.9. The values of χ^2 are shown in col.(4) of Table 2.4(b) and the P-values are indicated in col.(5) of the same. In only one case, the value reaches the 5% level of significance, but on the other hand, one value falls below the lower 1% point. The P-values are well-spread over the interval (0, 1), as they should be under the null hypothesis, viz., that the method of sampling makes no difference. The total of the fourteen χ^2's is 46.313, which is a very reasonable value for a χ^2 with 42 d.f. (P-value nearly 0.30). The one or two exceptional values of χ^2 may easily be explained by chance : The quantity in the numerator of χ^2 has only 3 degrees,

of freedom[1].

2.7.10. One may now attempt to interpret these results.

2.7.11. The χ^2-test for homogeneity presupposes that the samples are ordinary random. But the probability samples naturally have some element of clustering, all words on a sample line being selected together. It will be seen in Chapter 5 that there are small but positive autocorrelations between lengths of neighbouring words. The probability samples should therefore have slightly larger sampling errors than random samples of equal size; and the same holds for the subsamples of probability samples [Cochran, 1963, Chap.9; Sukhatme, 1954, Chap. VI]. This explains the small though nonsignificant upward bias of the χ^2's in col. (4) of Table 2.4(a)[2].

2.7.12. Consider now the χ^2's in col. (6), relating to the subsamples of the systematic samples. The small upward bias due to

---

1/ If in this test, the variance of $\bar{x}$ is estimated from the subsample averages of the probability sample, the ratio may be regarded as F with 3 and 3 d.f.'s. This ratio is found to be larger than 1 and also smaller than 1, significantly, in a number of cases. This is obviously due to the unreliable nature of such estimates of $V(\bar{x})$ depending on 3 d.f.'s only. The general picture remains unchanged, however. The same thing was seen in another way. When the variation between the subsample means of the probability sample was compared in the same manner with the more precise estimate of $V(\bar{x})$, the results were quite satisfactory, on the whole, but in individual cases the numerator based on 3 d.f. differed significantly from the estimate placed in the denominator.

2/ This indirect method of assessing the effect of autocorrelations is rather insensitive, for it cannot clearly show the effect of small intraline correlations by giving too high values of χ^2; by direct methods such correlations have been found to be definitely significant in Chapter 5.

the use of line-clusters seems to be present here also and in more or less the same degree, broadly speaking. One might say that the subsamples of the systematic samples are about as variable as probability samples of equal size, that is to say, they are slightly more variable than unrestricted random samples of words.

2.7.13. The same conclusion emerges from the χ^2-tests summarised in Table 2.4.(b). The subsamples of the systematic samples have similar sampling errors as probability samples of the same size.

2.7.14. It may also be said that, to a first approximation, the probability samples and the subsamples of systematic samples behave like unrestricted random samples of the same size.

2.7.15. Why then should there be some downward bias in the χ^2's of col.(8) of Table 2.4(a)? The downward bias is significant at 5% level. And it should be realised that two probability samples or two systematic samples from the same work would tend to give χ^2's on the high side, as already found. Why is it that the combined systematic sample agrees too closely with the combined probability sample, although the subsamples of the systematic sample are like probability samples ? A plausible explanation, a little speculative, is given below.

2.7.16. Suppose that the population of word-lengths is not perfectly in a random order, but that there are small patches, of length varying from, say, one paragraph to a few pages, which differ from one another in respect of the average or distribution

of word-length, but which are themselves relatively homogeneous.' This, it may be stated, is a realistic picture [vide Chapter 5], although the patches are not quite conspicuous in many works. Under such circumstances, a subsample of a systematic sample as used in the present study, may, because of the long interval between successive sample lines, miss many of the patches altogether; also different subsamples would tend to sample different sets of patches. This feature would tend to increase the between subsample differences for the systematic sample. But the combined systematic sample may sample most of the small patches of the work, and so the error of the combined systematic sample would not be appreciably increased by the between patches variation.

2.7.17. If now one remembers that 'systematic' sampling ensures a more even spread of sample lines than any probability sample can do, that is to say, the systematic samples are in a certain sense stratified samples of words, one can understand why the combined systematic samples agree too closely with the combined probability samples. In the case of subsamples of the systematic samples, this effect of stratification may have counteracted the effect of between patches variation. Combined systematic samples seem to be less variable than the subsamples of systematic samples indicate, that is, less variable than probability samples of the same size.[1/]

1/ For works numbered 30-35 in Table 2.4(a) the sampling fraction for systematic sampling was much higher than for other works. But it is not clear whether this made any appreciable difference in the properties of systematic sample estimates discussed here. However, these works are very short and hence relatively homogeneous.

2.7.18. Anyway, for practical purposes, the following conclusions will be used. There is, first, the broad conclusion that systematic sample estimates behave like probability sample estimates to a close degree of approximation. More precisely, we may state :

(1) The combined systematic sample is probably having a slightly smaller sampling error than a probability sample of the same size; and

(2) The subsamples of the systematic sample behave like probability samples of the same size, and hence overestimate to a small extent, the error of the combined systematic sample.

2.8.1. Complete Count Figures for Poems : All the poems of Tagore listed in Table 3.2 of Chapter 3 were rather short and therefore counted completely, excepting one very long poem, viz.,"Puraskar" where every eighth stanza starting from the second stanza was chosen.

2.8.2. Theoretically, the question of sampling errors may not be completely answered even by counting all words or syllables of any work in prose or poetry. The whole work may be conceived of as one sample from a parent population, or better perhaps, as one realization of a stochastic process, and one may be interested in drawing inferences about the underlying population or stochastic process. The problem posed here is exceedingly complex, and no attempt has been made in this study to attack this general problem.

2.8.3. The following crude method was used to get very rough ideas about the reliabilities of complete count data for poems, and also of the data for the poem, "Puraskar" : Each poem or poetical piece was split into two parts by dividing it suitably near the middle[1]. Averages of word-length and certain other characteristics were computed separately for the two parts, and, of course, for the whole piece [vide Table 3.2]. The divergence between the two part figures was taken as a rough indicator of reliability.

1/ This was easy in a majority of cases. For poems comprising a number of stanzas, the two parts may get exactly or nearly equal number of stanzas. In some other cases, again, the poems appeared to be naturally divided into several parts, the theme changing from one to the next; in such cases, these natural divisions were suitably used. In a few remaining cases, however, the formation of parts had to be more or less arbitrary.

Chapter 3: Word-length in Bengali — A Broad Survey[1/]

3.0 The object of the present chapter is to make a broad study of the word-length distributions, averages etc., estimated for different works in Bengali prose, mostly fiction, and also for a few representative poems.[2/] This would give some dimensional ideas about word-length in different types of works and about historical trends, if any. One may also consider the within author variation in word-length, that is to say, how far word-length distributions can indicate an author's individual style.

3.1.1. The Material : Tables 3.1 to 3.4 present the word-length data collected for different works. Word-length has been measured by the number of syllables.[3/]

1/ See Appendix 3 for some small-scale investigations on English.

2/ The figures for poems are mostly based on complete counts, while those for prose works are derived from probability and/or systematic samples of words.

3/ Vide Appendix 1 for some studies on word-length in terms of letters.

Table 3.1: Averages, standard deviations etc., of word-length in syllables, estimated for selected works in Bengali prose, separately by type of sample, and by subsamples (for average only).

author	work	type of sample	no. of sample lines	no. of sample words	average word-length in syllables by subsamples ss 1	ss 2	ss 3	ss 4	comb.	s.e. of combined average	standard deviation	entropy
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Vidyasagar	*Shakuntala	prob.	100	696	2.841	2.475	2.866	2.658	2.704	0.03921	1.1032	2.126
	*Sitar Vanavas	prob.	100	750	2.632	2.665	2.668	2.821	2.695	0.04527	1.2182	2.214
Bamkimchandra	*Durgeshnandini	prob.	100	577	2.682	2.463	2.814	2.392	2.579	0.05989	1.1268	2.120
		syst.	316	1782	2.593	2.593	2.592	2.585	2.591		1.0946	2.088
		pooled	416	2359	2.614	2.559	2.644	2.537	2.588	0.02962	1.1025	2.097
	Kapalkundala	syst.	90	493	2.788	2.600	2.662	2.556	2.645		1.2293	2.227
	*Visavriksha	prob.	99	611	2.404	2.534	2.531	2.419	2.470	0.04834	1.0567	2.042
		syst.	300	1852	2.457	2.469	2.500	2.397	2.455		1.0712	2.029
		pooled	399	2463	2.445	2.485	2.508	2.403	2.459	0.02408	1.0677	2.034
	Krishnakanter Will	prob.	200	1777	2.330	2.281	2.366	2.378	2.340	0.02801	1.0100	1.966
		syst.	128	749	2.370	2.392	2.415	2.318	2.372		1.0796	2.037
		pooled	328	2526	2.342	2.316	2.379	2.360	2.350	0.02349	1.0312	1.989
	Anandamath	prob.	200	1109	2.395	2.442	2.496	2.440	2.443	0.03130	1.0199	1.963
		syst.	133	801	2.508	2.510	2.419	2.245	2.438		1.0512	2.034
		pooled	333	1910	2.441	2.470	2.465	2.352	2.441	0.02385	1.0332	2.016
	Devi Chaudhurani	prob.	200	1174	2.353	2.219	2.189	2.378	2.283	0.02980	0.9276	1.862
		syst.	146	833	2.320	2.251	2.090	2.269	2.227		0.8395	1.838
		pooled	346	2007	2.339	2.232	2.147	2.333	2.260	0.02279	0.8925	1.855
	Rajsinha	prob.	250	1423	2.545	2.467	2.434	2.486	2.482	0.02788	1.0782	2.067
		syst.	96	507	2.715	2.634	2.512	2.520	2.596		1.1733	2.137
		pooled	346	1930	2.593	2.509	2.455	2.495	2.512	0.02394	1.1040	2.089

Table 3.1 (contd.)

author	work	type of sample	no. of sample lines	no. of sample words	average word-length in syllables by subsamples ss 1	ss 2	ss 3	ss 4	comb.	s.e. of combined average	standard deviation	entropy
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Rabindranath	Bouthakuranir Hat	prob.	200	1592	2.358	2.360	2.409	2.421	2.386	0.02677	0.9705	1.933
		syst.	100	827	2.498	2.392	2.266	2.484	2.406		0.9659	1.924
		pooled	300	2419	2.406	2.371	2.359	2.442	2.393	0.02172	0.9690	1.931
	Rajarsi	prob.	200	1632	2.394	2.419	2.432	2.449	2.424	0.02184	0.9909	1.941
		syst.	88	689	2.454	2.546	2.358	2.513	2.467		0.9575	1.922
		pooled	288	2321	2.412	2.458	2.410	2.467	2.437	0.01832	0.9813	1.937
	Chokher Bali	syst.	156	1318	2.329	2.458	2.366	2.322	2.366		0.9102	1.862
	*Gora	prob.	100	889	2.291	2.353	2.417	2.268	2.331	0.03619	0.9054	1.838
		syst.	203	1824	2.360	2.374	2.359	2.283	2.345		0.9471	1.883
		pooled	303	2713	2.339	2.367	2.377	2.278	2.341	0.02072	0.9317	1.870
	Ghare Baire	prob.	200	1901	2.047	2.088	2.141	2.102	2.093	0.02106	0.8338	1.659
	Chaturanga	prob.	200	1458	2.452	2.261	2.254	2.349	2.326	0.02482	0.9150	1.847
		syst.	113	854	2.347	2.269	2.293	2.274	2.296		0.9054	1.850
		pooled	313	2312	2.411	2.264	2.268	2.323	2.315	0.01971	0.9116	1.849
	*Sheser Kavita	prob.	100	735	2.122	2.332	2.173	2.125	2.186	0.03824	0.9371	1.758
		syst.	181	1284	2.267	2.150	2.222	2.160	2.204		0.9097	1.769
		pooled	281	2019	2.212	2.213	2.204	2.147	2.198	0.02307	0.9198	1.769
	Yogayog	syst.	142	1187	2.130	2.137	2.252	2.157	2.168		0.9133	1.757
Pramatha Choudhury	*Char-Yari Katha	prob.	100	872	2.049	2.014	2.050	2.124	2.060	0.02728	0.8560	1.662
	Birbaler Halkhata	prob.	100	1041	2.249	2.355	2.314	2.331	2.311	0.04096	1.0761	1.999
Saratchandra	*Pallisamaj	prob.	100	890	2.225	2.171	2.265	2.192	2.212	0.03527	0.9282	1.817
	*Pather Dabi	prob.	100	815	2.291	2.300	2.183	2.133	2.228	0.03602	0.8901	1.821

Table 3.1 (contd.)

author	work	type of sample	no. of sample lines	no. of sample words	average word-length in syllables by subsamples ss 1	ss 2	ss 3	ss 4	comb.	s.e.of combined average	standard deviation	entropy
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Bibhutibhusan	*Pather Panchali	prob.	100	922	2.284	2.310	2.206	2.202	2.250	0.03336	0.9134	1.830
		syst.	172	1630	2.291	2.211	2.302	2.320	2.279		0.9013	1.824
		pooled	272	2552	2.289	2.246	2.266	2.275	2.269	0.02005	0.9057	1.829
	Aparajita	syst.	201	1894	2.225	2.248	2.296	2.333	2.273		0.9413	1.855
	*Devayan	prob.	100	931	2.172	2.140	2.059	2.134	2.126	0.03255	0.8573	1.702
		syst.	244	2245	2.107	2.172	2.131	2.160	2.142		0.8507	1.701
		pooled	344	3176	2.126	2.162	2.110	2.152	2.138	0.01762	0.8526	1.702
Jajabar	*Dristipat	prob.	100	772	2.394	2.412	2.367	2.382	2.389	0.03954	1.0290	1.977
		syst.	213	1591	2.400	2.425	2.293	2.472	2.398		1.0413	1.986
		pooled	313	2363	2.398	2.421	2.318	2.443	2.395	0.02260	1.0374	1.988
	*Janantik	prob.	100	690	2.368	2.098	2.357	2.364	2.293	0.04210	0.9423	1.831
Muztaba Ali	*Chacha Kahini	prob.	100	778	2.222	2.234	2.092	2.206	2.189	0.03527	0.8418	1.691
	*Deshe Videshe	prob.	100	791	2.139	2.214	2.218	2.215	2.172	0.03159	0.8671	1.720
Bamkimchandra	Samya	syst.	114	1010	2.429	2.605	2.686	2.788	2.619		1.2516	
Rabindranath	Bamkimchandra	syst.	139	1237	2.767	2.730	2.590	2.642	2.682		1.1620	
	Vishwavidyalay	syst.	103	1009	2.339	2.271	2.458	2.296	2.339		0.9876	
	Kabuliwalla	syst.	86	779	2.439	2.503	2.539	2.523	2.501		1.0240	
	Kshudhita Pasan	syst.	125	1192	2.456	2.601	2.505	2.537	2.524		1.0537	
	Laboratory	syst.	131	1228	2.192	2.106	2.115	2.108	2.131		0.8860	

Table 3.2: Distribution of words by length in syllables, estimated for selected works in Bengali prose, separately by type of sample and by subsamples

author	work	type of sample	sub-sample	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9 a/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Vidyasagar	Shakuntala	prob.	1	170	8.82	30.00	39.42	15.29	2.94	3.53			
			2	181	17.68	38.67	27.08	11.60	4.97				
			3	164	7.93	33.54	34.14	16.46	4.88	2.44	0.61		
			4	181	11.60	35.91	34.82	11.60	4.97	2.10			
			comb.	696	11.64	34.63	33.76	13.65	4.45	1.73	0.14		
	Sitar Vanavas	prob.	1	193	15.03	33.16	33.16	12.95	4.14	1.04	0.52		
			2	191	16.23	31.41	35.09	7.85	7.33	1.05	0.52	0.52	
			3	187	16.04	30.49	34.76	13.37	2.14	2.67			0.53(10)
			4	179	12.85	30.17	34.07	11.73	8.38	1.68	1.12		
			comb.	750	15.07	31.33	34.27	11.47	5.47	1.60	0.53	0.13	0.13(10)
Bamkim-chandra	Durgesh-nandini	prob.	1	148	12.16	32.44	37.16	12.84	4.05	1.35			
			2	147	17.69	36.74	31.97	10.20	2.04	1.36			
			3	129	10.85	32.56	36.43	10.85	6.20	0.78	1.55	0.78	
			4	153	20.26	39.86	27.45	6.54	4.58	1.31			
			comb.	577	15.42	35.53	33.11	10.05	4.16	1.21	0.35	0.17	
		syst.	1	487	12.94	37.78	31.68	13.96	2.87	0.41		0.41	
			2	408	11.96	39.22	32.84	12.01	2.94	0.74	0.49		
			3	424	12.97	37.50	33.73	10.61	3.77	1.18	0.24		
			4	463	16.41	36.07	30.89	9.72	4.54	0.86	1.51		
			comb.	1782	13.58	37.60	32.21	11.82	3.54	0.79	0.56	0.11	
		pooled	comb.	2359	14.03	37.09	32.43	11.23	3.69	0.89	0.51	0.13	

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no. of sample words	percentage of words by length in syllables 1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Kapalkundala	syst.	1	104	7.69	43.28	25.00	13.46	7.69	2.88			
			2	135	18.52	37.04	24.44	10.37	6.67	2.22		0.74	
			3	130	13.85	33.07	34.62	10.77	6.92	0.77			
			4	124	21.77	32.26	27.42	9.68	5.64	2.42	0.81		
			comb.	493	15.82	36.11	27.99	10.95	6.69	2.03	0.20	0.20	
	Visavriksha	prob.	1	151	13.91	44.37	31.79	7.28	2.65				
			2	146	13.70	45.21	25.34	9.59	3.42	1.37	1.37		
			3	147	17.69	36.06	28.57	11.56	5.44	0.68			
			4	167	19.16	37.12	29.34	11.98	1.80	0.60			
			comb.	611	16.20	40.59	28.81	10.15	3.27	0.65	0.33		
		syst.	1	505	16.04	43.57	25.54	10.10	3.56	0.79	0.20	0.20	
			2	431	16.24	41.76	26.68	11.14	3.02	0.70	0.46		
			3	448	14.06	44.19	27.68	8.49	4.02	0.89	0.45		0.22
			4	468	16.67	43.16	28.85	7.48	2.99	0.64	0.21		
			comb.	1852	15.77	43.20	27.16	9.29	3.40	0.76	0.32	0.05	0.05
		pooled	comb.	2463	15.87	42.55	27.57	9.50	3.37	0.73	0.32	0.04	0.04

Table 3.2 (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables 1	2	3	4	5	6	7	8	9+ [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bankim-chandra	Krishnakenter Will	prob.	1	220	17.73	44.56	27.27	6.36	3.18	0.45	0.45		
			2	222	22.52	40.09	25.23	10.36	0.90	0.45	0.45		
			3	201	14.43	42.28	31.84	8.96	2.49				
			4	226	20.80	50.88	22.12	4.87	1.33				
			5	222	18.02	37.39	28.38	10.81	4.50	0.45			0.45
			6	234	23.50	41.46	25.64	7.26	1.71	0.43			
			7	217	21.20	42.40	25.80	6.45	3.23	0.46	0.46		
			8	235	14.89	42.98	30.22	7.23	4.68				
			comb.	1777	19.19	42.77	27.01	7.77	2.76	0.28	0.17	-	0.06
		syst.	1	192	20.31	40.63	28.12	6.25	3.65		0.52	0.52	
			2	194	22.16	34.02	30.93	8.76	3.61	0.52			
			3	171	22.22	38.01	25.15	8.77	3.51	1.17	1.17		
			4	192	18.23	43.23	29.69	6.25	2.60				
			comb.	749	20.69	38.99	28.57	7.48	3.34	0.40	0.40	0.13	
		pooled	comb.	2526	19.64	41.65	27.47	7.68	2.93	0.32	0.24	0.04	0.04

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Anandamath	prob.	1	144	15.97	41.67	30.56	9.03	2.08	0.69			
			2	142	17.60	45.08	23.95	10.56	2.11	-	0.70		
			3	144	16.67	38.20	31.25	11.11	1.39	0.69	0.69		
			4	150	14.00	44.67	30.00	9.33	1.33	0.67	-		
			5	132	13.64	45.45	31.06	6.06	2.27	1.52	-		
			6	138	17.39	32.61	34.06	10.87	2.90	1.45	0.72		
			7	135	19.26	34.82	27.41	14.81	2.96	-	0.74		
			8	124	15.32	42.74	32.26	8.87	0.81	-	-		
			comb.	1109	16.23	40.67	30.03	10.10	1.98	0.63	0.36		
		syst.	1	197	13.20	44.66	27.92	9.64	3.05	0.51	0.51	0.51	
			2	206	15.05	40.29	31.07	8.74	2.91	1.46	-	0.48	
			3	186	15.59	43.01	24.73	11.29	4.30	-	0.54	0.54	
			4	212	23.11	41.51	26.42	6.13	2.36	0.47			
			comb.	801	16.85	42.32	27.59	8.86	3.12	0.62	0.25	0.37	
		pooled	comb.	1910	16.49	41.36	29.01	9.58	2.46	0.63	0.31	0.16	

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Devi Chou-dhurani	prob.	1	136	16.15	44.62	26.92	9.23	2.31	0.77			
			2	147	13.73	49.01	30.72	4.58	1.96	-			
			3	143	25.00	45.84	20.83	6.94	1.39	-			
			4	140	18.70	41.01	34.53	3.60	1.44	0.72			
			5	163	25.15	46.02	22.09	4.90	1.23	-			0.61
			6	154	18.18	46.11	27.87	5.84	-	-			
			7	154	18.75	37.51	34.03	6.94	2.08	0.69			
			8	137	15.65	42.18	32.65	8.16	1.36	-			
			comb.	1174	18.99	44.12	28.88	6.22	1.45	0.26			0.09
		syst.	1	203	17.73	48.28	23.15	6.90	3.45	-	0.49		
			2	195	20.51	41.54	31.79	4.62	1.54				
			3	234	26.07	43.16	26.92	3.42	0.43				
			4	201	18.91	44.28	29.35	5.97	1.49				
			comb.	833	21.01	44.30	27.73	5.16	1.68		0.12		
		pooled	comb.	2007	19.83	44.20	28.40	5.78	1.54	0.15	0.05		0.05

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9+ a/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Rajsinha	prob.	1	131	17.56	35.88	29.77	10.69	5.34	-	0.76		
			2	139	14.39	43.16	25.18	13.67	2.16	0.72	0.72		
			3	146	11.64	44.52	31.52	8.90	3.42	-	-		
			4	145	17.92	41.38	28.28	8.97	2.76	0.69	0.71		
			5	141	10.64	46.10	25.52	10.64	4.26	2.13	0.71		
			6	147	16.33	38.09	30.62	12.24	1.36	1.36	-		
			7	139	22.30	42.45	20.30	10.79	1.44	0.72	-		
			8	141	12.77	41.84	26.24	12.06	4.96	1.42	0.71		
			9	148	18.92	34.45	31.76	9.46	4.73	0.68			
			10	146	21.92	34.26	29.45	8.90	4.79	0.68			
			comb.	1423	16.44	40.20	28.11	10.61	3.51	0.84	0.28		
		syst.	1	130	16.92	30.77	33.85	9.23	5.38	1.54	0.77	0.77	0.77(9)
			2	123	9.76	43.09	28.47	13.82	2.43	2.43			
			3	131	13.74	41.98	29.01	10.69	3.82	0.76			
			4	123	16.26	39.84	28.45	9.76	3.25	1.63	0.81		
			comb.	507	14.20	38.86	29.98	10.85	3.75	1.58	0.39	0.20	0.20
		pooled	comb.	1930	15.85	39.84	28.60	10.67	3.58	1.04	0.31	0.05	0.05

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no.of sample words	Percentage of words by length in syllables 1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Bouthakura-nir Hat	prob.	1	195	17.44	39.48	32.31	7.18	3.08	0.51			
			2	213	16.43	46.01	30.05	4.69	2.82	-			
			3	218	17.89	43.58	27.98	7.80	2.75	-			
			4	187	15.51	41.18	32.22	7.49	1.60	-			
			5	198	19.19	39.90	29.29	6.06	3.54	2.02			
			6	208	14.90	42.32	31.73	9.13	1.92	-			
			7	197	14.21	43.66	30.96	6.09	4.57	0.51			
			8	176	19.89	36.36	32.95	6.25	4.55	-			
			comb.	1592	16.90	41.71	31.09	6.85	3.08	0.38			
		syst.	1	211	11.85	42.17	35.55	5.69	4.27	0.47			
			2	212	16.04	40.56	33.96	7.55	1.42	0.47			
			3	218	23.39	37.62	31.19	5.05	2.29	0.46			
			4	186	15.59	35.48	38.71	6.45	2.69	1.08			
			comb.	827	16.81	39.06	34.70	6.17	2.66	0.60			
		pooled	comb.	2419	16.87	40.80	32.33	6.61	2.94	0.45			

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								a/
					1	2	3	4	5	6	7	8	9
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Rajarsi	prob.	1	189	15.87	41.80	30.15	7.94	2.65	1.06	-		0.53(10)
			2	197	18.78	43.66	26.90	8.63	1.52	-	0.51		
			3	218	13.30	46.79	30.74	5.96	2.75	-	0.46		
			4	202	16.83	36.63	35.64	7.92	2.48	0.50	-		
			5	211	18.48	42.18	27.49	8.06	2.84	0.95	-		
			6	210	11.90	45.25	30.00	8.57	3.33	0.95	-		
			7	197	13.20	41.62	32.49	10.66	1.52	-	0.51		
			8	208	11.54	48.08	29.81	8.65	0.96	0.96	-		
			comb.	1632	14.95	43.32	30.39	8.27	2.27	0.55	0.18	-	0.06(10)
		syst.	1	172	13.37	44.18	31.40	6.98	2.91	1.16			
			2	183	10.38	40.98	37.16	8.20	1.64	1.64			
			3	176	19.88	35.23	35.23	8.52	1.14				
			4	158	9.49	47.47	28.48	12.66	1.27	-	0.63		
			comb.	689	13.35	41.80	33.24	9.00	1.74	0.73	0.15		
		pooled	comb.	2321	14.48	42.87	31.24	8.49	2.11	0.60	0.17		0.04(10)

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9+ [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Chokher Bali	syst.	1	337	16.62	43.32	31.45	7.72	0.89				
			2	308	12.99	43.19	32.79	8.44	1.62	0.65	0.32		
			3	322	15.22	46.89	26.09	9.94	1.55	0.31			
			4	351	15.95	45.31	31.34	5.70	1.42	0.28			
			comb.	1318	15.25	44.69	30.42	7.89	1.37	0.30	0.08		
	Gora	prob.	1	213	15.02	48.36	30.52	4.69	1.41				
			2	221	13.12	50.68	26.70	6.79	2.71				
			3	216	18.06	40.28	27.31	10.65	3.70				
			4	239	15.90	50.63	26.78	4.60	1.67	0.42			
			comb.	889	15.52	47.59	27.78	6.64	2.36	0.11			
		syst.	1	492	16.46	44.51	28.46	7.93	2.44	0.20			
			2	439	14.81	46.47	28.93	6.83	2.28	0.68			
			3	473	13.53	50.32	27.27	5.71	2.54	0.21	0.21	0.21	
			4	420	18.81	48.57	20.24	10.24	2.14				
			comb.	1824	15.84	47.42	26.37	7.62	2.36	0.27	0.05	0.05	
		pooled	comb.	2713	15.74	47.48	26.83	7.30	2.36	0.22	0.04	0.04	

Table 3.2 (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Chaturanga	prob.	1	179	12.85	41.34	34.64	8.38	2.23		0.56		
			2	162	15.43	43.21	27.78	11.11	2.47				
			3	179	20.11	44.14	28.49	5.03	2.23				
			4	177	15.25	49.16	29.95	5.08		0.56			
			5	181	20.44	39.23	35.92	3.31	1.10				
			6	205	20.00	45.85	27.32	3.41	2.93		0.49		
			7	196	14.80	47.96	27.04	7.14	2.55	0.51			
			8	179	11.17	54.75	26.26	6.14	1.12		0.56		
			comb.	1458	16.32	45.75	29.63	6.10	1.85	0.14	0.21		
		syst.	1	216	14.81	46.31	30.09	6.94	1.85				
			2	219	18.26	46.58	27.85	5.02	1.83	0.46			
			3	215	17.67	49.30	21.86	8.84	1.86	0.47			
			4	204	20.10	39.71	33.82	5.39	0.98				
			comb.	854	17.68	45.55	28.34	6.56	1.64	0.23			
		pooled	comb.	2312	16.83	45.67	29.15	6.27	1.77	0.17	0.13		

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Ghare Baire	prob.	1	253	22.92	58.89	14.62	2.77	0.40	0.40			
			2	256	21.88	54.68	17.97	3.52	1.56	0.39			
			3	238	21.85	54.62	19.33	3.36	0.42		0.42		
			4	237	19.83	59.92	11.39	8.02	0.84				
			5	225	20.44	56.44	18.67	3.56	0.89				
			6	192	15.62	58.34	17.19	7.29	1.04	0.52			
			7	257	19.07	58.75	15.95	4.67	1.56				
			8	243	22.63	56.38	12.76	5.76	2.06	0.41			
			comb.	1901	20.67	57.23	15.94	4.79	1.10	0.21	0.05		
	Sheser Kavita	prob.	1	188	21.28	53.19	18.09	6.91	0.53				
			2	178	12.92	57.30	17.98	7.87	3.37	0.56			
			3	185	17.84	56.76	18.38	4.32	2.70				
			4	184	23.91	50.00	21.20	2.72	1.63				0.54(11)
			comb.	735	19.05	54.28	18.91	5.44	2.04	0.14	-	-	0.14(11)
		syst.	1	307	14.33	54.07	23.45	7.17	0.65	0.33	-	-	
			2	340	20.88	53.24	18.53	5.29	1.47	0.59	-	-	
			3	311	20.90	48.87	20.26	6.75	1.61	1.29	-	0.32	
			4	326	16.56	59.51	17.48	5.21	0.61	0.31	0.31	-	
			comb.	1284	18.22	53.97	19.86	6.07	1.09	0.62	0.08	0.08	
		pooled	comb.	2019	18.52	54.09	19.51	5.84	1.44	0.45	0.05	0.05	0.05(11)

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Yogayog	syst.	1	293	19.45	57.34	16.72	4.10	2.05	0.34			
			2	307	20.85	55.70	14.66	6.51	2.28				
			3	294	17.69	53.74	17.35	9.18	1.02	1.02			
			4	293	20.48	55.28	15.36	6.83	1.37	0.34	0.34		
			comb.	1187	19.63	55.52	16.01	6.66	1.68	0.42	0.08		
Pramatha Choudhury	Char-Yari Katha	prob.	1	224	20.54	62.06	12.05	3.57	0.89	0.89			
			2	213	23.94	56.81	14.08	4.23	0.94				
			3	218	22.48	59.17	11.01	5.96	0.92	0.46			
			4	217	23.96	49.31	19.82	5.53	0.46	0.46	0.46		
			comb.	872	22.71	56.88	14.22	4.82	0.80	0.46	0.11		
	Birbaler Halkhata	prob.	1	277	23.47	42.24	23.82	6.86	3.61				
			2	262	20.99	42.75	24.43	5.73	4.58	0.76	0.76		
			3	242	21.90	42.57	24.79	7.44	1.24	1.24	0.41	-	0.41
			4	260	18.08	47.70	22.69	7.69	2.69	0.77	0.38		
			comb.	1041	21.13	43.81	23.92	6.92	3.07	0.67	0.38	-	0.10
Sarat-chandra	Pallisamaj	prob.	1	218	20.18	48.17	25.23	2.75	2.75	0.92			
			2	228	18.42	51.31	25.88	3.51	0.88				
			3	215	19.07	48.83	24.65	3.72	2.79		0.47	0.47	
			4	229	23.14	44.55	24.45	5.68	2.18				
			comb.	890	20.23	48.20	25.06	3.93	2.14	0.22	0.11	0.11	

Table 3.2: (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Sarat-chandra	Pather Dabi	prob.	1	220	10.09	40.92	33.18	5.45	1.36				
			2	197	16.75	49.23	24.37	8.12	0.51	0.51	0.51		
			3	202	22.77	44.55	25.25	6.44	0.99				
			4	196	22.96	44.90	29.08	2.04	1.02				
			comb.	815	20.37	44.79	28.10	5.52	0.98	0.12	0.12		
Bibhuti-bhusan	Pather Panchali	prob.	1	222	17.57	45.04	30.18	5.86	1.35				
			2	239	16.32	50.62	21.76	8.37	2.93				
			3	233	20.17	49.35	22.32	6.01	2.15				
			4	228	20.18	51.31	20.18	4.82	3.51				
			comb.	922	18.55	49.13	23.54	6.29	2.49				
		syst.	1	422	18.25	47.16	25.59	6.16	2.13	0.47	0.24		
			2	436	18.58	50.46	23.39	6.42	1.15				
			3	397	15.62	49.12	25.94	8.06	1.26				
			4	375	14.93	49.34	27.73	5.60	1.60	0.80			
			comb.	1630	16.93	49.02	25.58	6.56	1.53	0.31	0.06		
		pooled comb.		2552	17.52	49.06	24.84	6.47	1.88	0.20	0.04		

Table 3.2 (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bibhuti-bhusan	Aparajita	syst.	1	511	18.98	48.93	25.44	5.28	0.78	0.39			0.20(10)
			2	484	17.77	48.76	26.03	5.99	1.24	0.21			
			3	452	18.36	46.91	25.88	4.87	3.54	0.44			
			4	447	18.12	45.19	26.40	6.49	3.36	0.22	0.22		
			comb.	1894	18.32	47.52	25.92	5.65	2.16	0.32	0.05	-	0.05(10)
	Devayan	prob.	1	233	22.75	50.63	15.45	9.01	2.16				
			2	236	19.07	55.08	19.07	6.36	0.42				
			3	239	20.50	60.67	12.97	4.60	0.84	0.42			
			4	223	18.83	57.86	16.59	4.48	2.24				
			comb.	931	20.30	56.07	16.00	6.12	1.40	0.11			
		syst.	1	562	21.00	56.58	15.30	5.34	1.42	0.36			
			2	559	17.53	59.04	17.17	5.37	0.89				
			3	574	19.69	53.83	17.60	7.49	1.22	0.17			
			4	550	19.09	54.19	20.18	5.27	0.91	0.18	0.18		
			comb.	2245	19.33	55.90	17.55	5.88	1.11	0.18	0.04		
		pooled	comb.	3176	19.62	55.95	17.10	5.95	1.20	0.16	0.03		

Table 3.2:(Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables 1	2	3	4	5	6	7	8	9- a/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Jajabar	Dristipat	prob.	1	188	15.96	44.68	26.06	10.64	2.66				
			2	194	14.95	46.91	25.77	8.76	2.58		1.03		
			3	199	17.59	46.73	19.60	13.57	2.51				
			4	191	20.42	40.31	26.18	8.38	4.19	-	-	0.52	
			comb.	772	17.23	44.69	24.35	10.36	2.98	-	0.26	0.13	
		syst.	1	402	15.42	47.27	24.63	8.70	2.49	1.49			
			2	409	16.63	44.01	25.43	9.29	3.67	0.73	0.24		
			3	386	17.35	51.56	20.47	7.25	2.33	0.78	-	0.26	
			4	394	16.24	42.13	25.13	12.18	3.30	1.02			
			comb.	1591	16.40	46.20	23.95	9.37	2.95	1.01	0.06	0.06	
		pooled	comb.	2363	16.67	45.70	24.08	9.69	2.96	0.68	0.13	0.08	
	Janantik	prob.	1	163	12.88	53.98	22.09	6.75	3.07	1.23			
			2	183	18.58	60.11	15.30	4.92	1.09				
			3	171	16.96	46.78	23.39	10.53	1.17	1.17			
			4	173	15.03	49.71	23.12	8.67	2.89	0.58			
			comb.	690	15.94	52.75	20.87	7.68	2.03	0.73			
Muztaba Ali	Chacha-Kahini	prob.	1	189	16.40	55.02	20.11	6.88	1.59				
			2	205	14.15	55.61	23.41	6.34	0.49				
			3	195	17.95	61.02	15.90	4.10	1.03				
			4	185	18.52	53.44	21.69	3.17	2.12	0.53	0.53		
			comb.	778	16.71	56.29	20.31	5.14	1.29	0.13	0.13		

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9- a/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Muztaba Ali	Deshe Videshe	prob.	1	208	19.23	56.25	17.79	4.81	1.92				
			2	182	15.38	56.60	20.33	6.59	1.10				
			3	193	17.62	53.89	19.17	7.77	1.55				
			4	208	21.15	55.78	15.38	5.77	1.44	-	0.48		
			comb.	791	18.46	55.62	18.08	6.19	1.52	-	0.13		
Bamkim-chandra	Samya	syst.	1	275	20.36	41.46	20.73	12.00	4.73	0.36			0.36(10)
			2	253	14.62	39.92	26.09	12.65	4.74	0.40	1.58		
			3	255	16.47	37.26	22.75	14.51	5.49	1.96	0.78	0.39	0.39
			4	227	10.13	38.77	27.31	14.10	5.73	3.52	0.44		
			comb.	1010	15.64	39.40	24.06	13.27	5.15	1.49	0.69	0.10	0.20(9,10)
Rabindra-nath	Bamkim-chandra	syst.	1	318	7.55	40.57	31.13	11.32	7.86	0.94	0.63		
			2	293	11.95	39.93	28.33	9.56	5.80	3.07	0.68	0.34	0.34
			3	310	13.23	40.65	25.48	16.13	3.87	0.32	0.32		
			4	316	13.29	36.08	32.59	10.76	6.01	0.95	0.32		
			comb.	1237	11.48	39.29	29.43	11.96	5.90	1.29	0.49	0.08	0.08
	Vishwa-vidyalay	syst.	1	254	16.14	54.33	16.54	7.48	3.94	1.18	0.39		
			2	255	17.65	49.40	24.71	5.88	1.18	1.18			
			3	240	13.75	45.84	25.83	10.83	2.92	0.83			
			4	260	16.92	49.23	23.08	8.85	1.92				
			comb.	1009	16.15	49.75	22.50	8.23	2.48	0.79	0.10		

Table 3.2 : (Contd.)

author	work	type of sample	sub-sample	no.of sample words	percentage of words by length in syllables									
					1	2	3	4	5	6	7	8	9	a/
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	
Rabindra-nath	Kabuliwalla	syst.	1	198	14.64	41.41	32.83	8.08	2.53	0.51				
			2	195	16.41	38.98	30.26	9.74	2.05	2.05	0.51			
			3	191	11.52	43.45	31.94	8.38	2.62	1.57	0.52			
			4	195	11.79	41.54	33.33	10.26	2.05	1.03				
			comb.	779	13.61	41.34	32.09	9.11	2.31	1.28	0.26			
	Kshudhita Pasan	syst.	1	307	13.68	40.71	35.18	8.47	1.30	0.33	-	0.33		
			2	288	9.03	46.17	29.17	10.42	3.47	1.04	0.35			0.35(10)
			3	297	13.13	43.78	30.30	8.08	3.03	1.01	-	0.67		
			4	300	11.67	44.68	29.33	9.33	3.33	1.33	-	0.33		
			comb.	1192	11.91	43.78	31.04	9.06	2.77	0.92	0.09	0.34		0.09(10)
	Laboratory	syst.	1	318	17.30	58.18	15.41	7.23	1.26	0.31	0.31			
			2	283	22.97	53.00	16.61	5.30	2.12					
			3	321	18.38	59.19	17.76	2.49	1.56	0.62				
			4	306	24.18	50.33	18.30	5.88	0.65	0.33	0.33			
			comb.	1228	20.60	55.29	17.02	5.21	1.38	0.33	0.16			

a/ The numbers inside brackets show the actual lengths of words falling in this length-class.

Table 3.3: Averages and standard deviations of word-length in syllables, for selected poems/poetry pieces in Bengali*

author	work	piece/poem	year of composition	no.of words by parts			average word-length in syllables			s.d. of word-length (syllables)
				1	2	comb.	part 1	part 2	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Michael M. Dutta	Meghanadabadha Kavya	first 200 lines of canto I	1861	523	458	981	2.55	2.59	2.57	1.225
Rabindranath Tagore	Prabhat Sangeet	Nirjharer Swapnabhanga	1882	119	102	221	2.15	2.08	2.11	0.833
	Manasee	Badhu	1888	203	206	409	2.22	2.04	2.13	0.830
		Meghadut	1890	244	350	594	2.77	2.90	2.85	1.426
	Sonar Taree	Sonar Taree	1892	83	88	171	2.27	1.98	2.12	0.763
		Puraskar	1893	160	154	314	2.29	2.68	2.48	1.026
		Niruddesh Yatra	1893	146	146	292	2.35	2.36	2.35	0.952
	Chitra	Sandhya	1894	121	158	279	2.45	2.47	2.46	0.975
		Urvashee	1895	160	165	325	2.91	2.81	2.86	1.279
	Kalpana	Varsamangal	1897	81	125	206	3.64	3.16	3.35	1.572
		Swapna	1897	125	111	236	2.90	2.45	2.69	1.183
	Kshanika	Krishnakali	1900	75	115	190	2.03	2.00	2.01	0.754
	Shishu	Virpurus	1903	157	169	326	203	1.88	1.95	0.653
	Balaka	Shajahan	1914	356	219	575	2.44	2.30	2.39	1.197
		Balaka	1915	94	174	268	2.74	2.33	2.47	1.056
	Shishu Bholanath	Khelabhola	1921	107	113	220	2.05	1.96	2.00	0.775
	Puravi	Satyendranath Dutta	1922	340	297	637	2.41	2.56	2.48	1.171
		Tapobhanga	1923	278	279	557	2.63	2.67	2.65	1.215
	Parishesh	Pranam	1931	109	135	244	2.67	2.84	2.76	1.400
		Bnashee	1932	133	179	312	2.30	2.34	2.32	1.041
	Shyamalee	Ami	1936	128	137	265	2.07	2.29	2.18	0.873
	Samayik Patra	Africa	1937	78	136	214	2.73	2.46	2.56	0.944
	Arogya	Ora Kaj Kare	1941	94	105	199	2.47	2.48	2.47	1.405

* A systematic sample of stanzas was examined for "Puraskar"; all other pieces were counted complete.

Table 3.4: Distribution of words by length in syllables for selected poems/poetry pieces in Bengali [a].

author	work	piece/poem	no.of words	percentage of words by length in syllables								
				1	2	3	4	5	6	7	8	9- [b]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Michael M. Dutta	Meghanadabadha Kavya	first 200 lines of Canto I	981	14.9	41.7	26.6	10.2	2.8	2.6	0.7	0.4	
Rabindranath Tagore	Prabhat Sangeet	Nirjharer Swapnabhanga	221	24.0	45.7	27.2	1.8	1.4				
	Manasee	Badhu	409	18.8	57.2	17.6	5.4	0.7	0.2			
		Meghadut	594	9.3	42.1	22.9	16.0	4.9	2.7	0.7	1.2	0.3(9,14)
	Sonar Taree	Sonar Taree	171	17.5	59.0	18.1	4.7	0.6				
		Puraskar	314	11.5	48.1	28.3	8.0	2.2	1.3	0.6		
		Nirudddesh Yatra	292	12.0	56.8	20.2	6.5	3.8	0.7			
	Chitra	Urvashee	325	9.2	39.7	22.8	17.8	6.5	3.1	0.6	0.3	
		Sandhya	279	10.4	51.3	25.1	10.4	1.4	1.1	0.4		
	Kalpana	Varsamangal	206	3.4	29.6	38.3	4.4	14.1	4.4	3.9	1.5	0.5(9)
		Swapna	236	8.9	46.2	23.3	14.8	3.0	3.4	-	0.4	
	Kshanika	Krishnakali	190	19.0	67.9	7.9	4.2	0.5	0.5			
	Shishu	Virpurus	326	19.0	71.5	5.2	4.0	0.3				
	Balaka	Shajahan	575	21.2	41.7	23.6	8.0	2.6	1.9	0.4	0.5	
		Balaka	268	12.0	48.9	25.0	11.2	1.5	0.8	-	0.8	
	Shishu Bholanath	Khelabhola	220	21.8	64.1	6.8	6.8	0.4				
	Puravi	Satyendranath Dutta	637	11.8	53.1	20.4	9.6	3.3	0.9	0.3	0.2	0.5(9,12)
		Tapobhanga	557	9.0	47.4	25.5	12.4	1.8	2.3	0.7	0.7	0.2(9)
	Parishesh	Pranam	244	9.4	44.3	25.4	13.5	2.0	2.9	0.4	1.2	0.8(9,10)
		Bnashee	312	16.0	52.2	21.5	6.7	1.6	1.3	0.3	0.3	
	Shyamalee	Ami	265	18.5	55.1	17.4	7.6	1.5				
	Samayik Patra	Africa	214	7.0	49.1	30.8	8.4	3.7	0.9			
	Arogya	Ora Kaj Kare	199	16.1	49.2	16.1	11.6	4.0	3.0			

[a] A systematic sample of stanzas was examined for "Puraskar"; all other pieces were counted complete.

[b] The number inside brackets show the actual lengths of words falling in this length-class.

3.1.2. Table 3.1 relates to the works in Bengali prose, fiction, essays and short stories. The works sampled are listed in cols. (1) and (2), and the sample sizes are shown in cols. (4) and (5). The average of word-length is given for each work, for each of the four independent and interpenetrating subsamples, and also for the combined sample [cols. (6) - (10)]. For each work is also presented the standard deviation of word-length [col.(12)] as well as the entropy $H = -\sum_r p_r \log_2 p_r$, where p_r is the proportion of r-syllabled words [col. (13)]. This latter measure has been used by Fucks (1952, 1954, 1955) and more extensively by Herdan (1956). Where both types of sampling have been used, the averages, s.d. and entropy are presented for each type of sample and also for the 'pooled' (probability plus systematic) sample. The type of sample is indicated in col. (3).

3.1.3. The standard error of the combined (all-subsample) mean is presented in col (11). For the probability sample from each work, this was calculated by using the expression given in para 2.4.2 of Chap.2. No such estimate is shown for the averages from systematic samples; but one might assume that these could be obtained approximately from the corresponding standard errors for the probability sample, by multiplying the latter by $\sqrt{\frac{n}{n'}}$, where n and n' are respectively the number of words in the probability and the systematic samples. This type of procedure was actually adopted for estimating the standard error ($S_{\bar{x}}$) for the 'pooled' average: The $S_{\bar{x}}$ for the pooled average was taken as $\sqrt{\frac{n}{n+n'}}$ times that for the probability sample. In view of the

findings of Chapter 2, Section 2.7, this should be sufficiently accurate for practical work. One may also use the subsample averages for making inferences about the true average.

3.1.4. Table 3.2 presents the word-length distributions for the different works in Bengali prose, that is to say, the estimates p_r of r-syllabled words (r = 1,2,...). The estimates are given by type of sample and by subsamples. The number of subsamples is sometimes larger than 4, in this table, for the probability samples. Combined (i.e., all-subsample) and pooled (i.e., probability plus systematic) distributions are also shown in every case. One may use this table for drawing inferences about the p_r's. Since the discussion has been largely confined to the averages of word-length, Table 3.1 practically suffices for the present purposes.

3.1.5. Tables 3.3 and 3.4 relate to Bengali poetry. Table 3.3 shows the average and s.d. of word-length for each poem/piece, while Table 3.4 shows the corresponding distributions of words by length in syllables. All the poems, excepting 'Puraskar', were completely counted; nevertheless, partwise averages are shown in Table 3.3 to throw some light on the reliabilities of the complete count averages [vide Section 2.8 of Chap. 2].

3.1.6. The coverage of works, especially essays, short stories and poetry, is admittedly small. The objectives are therefore very modest. One is to obtain a broad, dimensional, idea about word-length in different fields of Bengali literature. A second is to examine the historical

trends in word-length in different fields.

3.1.7. In spite of small coverage of works, many important conclusions can be read from Tables 3.1 to 3.4. For the sake of convenience, we shall first study the time-trend in word-length. The discussion will be mostly confined to the averages ($\bar{x}$).

3.2.1. Historical trends in word-length : The findings reported in this section may seem to be generally known, for the phenomenon is perhaps the most important in the last 100 years' history of literary Bengali. Objective data should, nevertheless, be useful.

3.2.2. Since word-length distributions tend to vary with the field of literature [vide Section 3.3, infra], the study may be initially confined to fiction.

3.2.3. Bengali fiction started with the average word-length $\bar{x}$ at a high level : The averages are nearly 2.7 for the two works by Ishwar Chandra Vidyasagar, one of the fathers of Bengali prose, written in the chaste, i.e., Sanskritised, style ("Sadhu bhasa"). But the average ($\bar{x}$) gradually came down during Bamkimchandra's period and later during Rabindranath Tagore's; and both these writers played leading roles in the movement.

3.2.4. Bamkimchandra's writings represent the attempts to find a suitable form of Bengali prose. He started with $\bar{x}$ of the order of 2.6 — vide "Durgeshnandini" (a historical novel) and "Kapalkundala" (a romance), his first two novels published in 1865 and 1866 respectively. But he gradually lowered the level of $\bar{x}$. The average is 2.46

(± 0.024) in "Visavriksha" (1873) and only 2.35 (± 0.023) in "Krishnakanter Will" (1878), his two famous novels on social problems. Even for "Rajsinha"[1/] (1882, revised 1893), a historical novel, the value is 2.51 (± 0.024); and "Anandamath" (1882) shows 2.44 (± 0.024). A surprisingly low value, 2.26 (± 0.023), is seen for "Devi Chaudhurani" (1884), a novel apparently similar to "Anandamath". It is clear that the major part of the fall in $\bar{x}$ occurred in the Bamkimchandra age. Bamkimchandra generally used the chaste form throughout his novels, that is, even in the conversational passages; but excepting for the verbs and the pronouns he had almost used the colloquial style ("chalita bhasa") in the conversations in "Devi Chaudhurani". In fact, this was the trend in his writings. The present author found in "Devi Chaudhurani" and "Rajsinha" occasional uses of colloquial verb and pronoun forms, of course, in the conversations.

3.2.5. Tagore's writings reflect a continuation of the same trend. The levels represented by "Krishnakanter Will" and "Devi Chaudhurani" are accepted as standard in the beginning. Tagore's first two novels, "Rajarsi" (1887) and "Bouthakuranir Hat" (1883) were both historical and written in the chaste style, which was, however, not as elevated as that in Vidyasagar or Bamkimchandra (early phase). They were written when he was hardly 20. This was the age of Bamkimchandra; and the $\bar{x}$-values are 2.44 (± 0.018) and 2.39 (± 0.022) respectively. Tagore

1/ Vide, in this connection, Section 3.3, especially para 3.3.8.

used "Sadhu bhasa", the chaste style, in "Chokher Bali" also. This is a novel on social problems published in 1903 and $\bar{x}$ is about 2.37 in this case. "Gora", written around 1910, uses the colloquial style in conversations, but this novel includes elevated discussions on social-religious problems, and $\bar{x}$ is 2.34 ($\pm$ 0.021), not much lower than in "Chokher Bali". The later novels of Tagore, "Chaturanga", "Ghare Baire" and others, are written in very unusual styles: They are often described as being too sketchy and not having some typical features of a novel. Anyway, "Chaturanga" (1916) was written in the chaste style throughout, and its mean word-length is 2.32 ($\pm$0.020) syllables. But "Ghare Baire", written in the same year, employed the colloquial style throughout and the mean is surprisingly low, only 2.09 ($\pm$ 0.021). "Sheser Kavita" and "Yogayog", written near 1929 in the colloquial style, also show low means, 2.20 ($\pm$ 0.023) and 2.17 respectively.

3.2.6. The figures for the post-Tagore writers hardly show any further trend. They seem to be choosing their levels of $\bar{x}$ from the range employed by Tagore, viz., 2.1 to 2.4. Colloquial style is generally employed now-a-days even for the non-conversational matter in the novels. Highly Sanskritised writing of Vidyasagar and Bamkimchandra (early period) has practically gone out of use. Modern Bengali novels have the effective range of $\bar{x}$ as 2.1 to 2.4, but very often the range is from 2.15 to 2.30, and the variation from author to author or work to work does not seem to be appreciable. Also, as we shall see, a part of such variation is due to varying weightage of conversational matter.[1/]

1/ Vide Chapter 4 for differences between lengths of conversational and other words.

3.2.7. The few figures for essays are in keeping with the above. "Samya" by Bamkimchandra shows $\bar{x}$ = 2.62; "Bamkimchandra" by Tagore, written on the death of Bamkimchandra in 1899 shows $\bar{x}$ = 2.68. Both are written in the chaste style, in an elevated language. But "Vishwavidyalay" by Tagore, a very serious literary effort, was composed in 1933 in the colloquial style, and has an $\bar{x}$ nearly 2.34. "Birbaler Halkhata", a collection of essays by Pramatha Chaudhuri, is also written in the same style; and the mean is 2.31.

3.2.8. A similar difference can be seen among the short stories: "Laboratory", composed in 1940, in the colloquial style, uses much shorter words, on the average, than "Kabuliwalla" or "Kshudhita Pasan" written earlier in the chaste style; the three averages are respectively 2.13, 2.50 and 2.52. However, this difference has been exaggerated by the much higher proportion of conversational matter in "Laboratory". "Chacha-Kahini", a collection of short stories by Muztaba Ali, published in 1952, . shows $\bar{x}$ = 2.19 ($\pm$ 0.035); this also is written in the colloquial style.

3.2.9. Although the coverage of essays and short stories is small, the findings seem to be generally true; for it is a matter of common knowledge that in all fields of Bengali prose the older chaste style employing compounds and other long words has been gradually replaced by the colloquial style over the last century of Bengali prose.

3.2.10. The data on poetry tell a different story. There is little evidence of any time-trend in the $\bar{x}$-values given in Table 3.3 for the different poems of Tagore. This presents a sharp contrast with what has been seen for Tagore's novels and also for his essays and short stories. The matter will be clearer in the following section. Bengali poetry seems to have behaved in a different way from Bengali prose so far as word-length is concerned.

3.3.1. Word-length in different types of literary works: It would be useful at this stage to have a close look at the $\bar{x}$-values for the different poems. The first point which strikes us from Table 3.3 is the remarkable variation in the average word-length $\bar{x}$ from one poem to another. The highest average is 3.35 syllables, for "Varshamangal", a lyric abounding with compounds; the part averages are 3.64 and 3.16.[1] The longest word found is of 14 syllables in "Meghadut", where $\bar{x}$=2.85 [vide last col. of Table 3.4]. "Urvashee" does not show any word with more than 8 syllables, but its average is also 2.86. Both these poems are on elevated topics. At the other end of the scale, one finds three poems with $\bar{x}$ below 2.1 — "Krishnakali", a lyric in light vein, with $\bar{x}$= 2.01, and the two poems for children, viz., "Virpurus" ($\bar{x}$ = 1.95) and "Khelabhola" ($\bar{x}$ = 2.00). The distributions in Table 3.4 reflect these differences.

1/ No other $\bar{x}$ above 3 has been found in the present study on Bengali novels, essays, short stories and poems, excepting for one short patch in "Visavriksha" by Bankimchandra [vide Chapter 5, Section 5.5].

3.3.2. .In between these two extremes, one finds almost continuous variation of the $\bar{x}$-values : "Pranam" — 2.76, "Swapna" — 2.69, "Tapobhanga" — 2.65, "Africa" — 2.56, "Puraskar" and "Satyendranath Dutta" — 2.48, "Balaka" and "Ora Kaj Kare" — 2.47, "Sandhya" — 2.46, "Shajahan" — 2.39, "Niruddesh Yatra" — 2.35, "Banshi" — 2.32, "Ami" — 2.18, "Badhu" — 2.13, "Sonar Taree" — 2.12 and "Nirjharer Swapnabhanga" — 2.11. These averages should be extremely interesting to those familiar with Tagore's poetry. It is important to notice that there is little time-trend in these values, and both high and low values occur equally among earlier and later poems of Tagore.

3.3.3. The extract from "Meghanadabadha Kavya" has $\bar{x}$ of the order of 2.55 or 2.6, which is not at all high for poetry. This contradicts popular impressions. This epic has a fair proportion of conversational matter; also the elevation of its style is largely due to the rich Sanskritised vocabulary, the meter (blank verse) and the imagery and rhetoric; and word-length seems to have very little to do with the elevation of style.

3.3.4. It is apparent that Bengali poems of today can have $\bar{x}$ anywhere from 2 to 2.9, roughly speaking, depending on theme, mood and meter. Values of $\bar{x}$ beyond this range are comparatively rare. The range of variation is much smaller for prose fiction, novels and short stories. The works of Vidyasagar and also Bamkimchandra (early phase) show $\bar{x}$ above 2.5 (upto 2.7) but such writing is now almost outmoded. In twentieth century fiction with conversational matter — or written

as speeches or thoughts of the leading character or characters — the effective range is from about 2.1 to 2.4; $\bar{x}$ above 2.5 would now appear to be artificial. The upper limit may be raised for fiction not including conversational matter is any appreciable proportion, like "Kabuliwalla" and "Kshudhita Pasan". For essays, where conversation does not appear, the effective range may be taken as 2.3 to 2.7 [consider "Birbaler Halkhata" and "Bamkimchandra"].

3.3.5. It may be pointed out here that a poem with a high value of $\bar{x}$, say 2.7, does not appear to be so unusual as an essay with the same value of $\bar{x}$. This is because long compounds and ornamentation seem to be natural in poetry, or to put it differently, the readers do not insist that poetry should employ everyday language to the same extent as they do for prose works. This explains why while $\bar{x}$ showed a falling trend in Bengali prose during the last 100 years, no such trend seems to be visible for the poetical literature of Bengali.

3.3.6. In continuation of the same point, it may be observed that the different poems in many poetical works by Tagore (say) often display conspicuous variation in the average value of word-length. ("Gitanjali" seems to be an important exception.) A common thread may be there running through all or most of the poems, but the styles of consecutive poems, composed within a few days from one another often vary markedly and erratically. The significance of word-length in Bengali poetry is very different from the significance of word-length in Bengali prose.

3.3.7. A high value of $\bar{x}$ indicates an elevated style using a high proportion of 'tatsama' words (Sanskrit words in unmodified form) and compounds, while a low $\bar{x}$ is generally associated with a high proportion of 'tatbhava' words (i.e., Prakrit words, i.e., Sanskrit words in modified form). Whether the verbs and pronouns have the 'Sadhu' (chaste) or the 'chalita' (colloquial) forms seems to be of minor direct consequence in poetry and the colloquial form may be used in elevated poetry without any jarring effect on the ear. For instance, 'chalita' verbs are used in "Meghadut" ($\bar{x}$ = 2.85) with an elevated style, while "Badhu" and "Nirjharer Swapnabhanga" with $\bar{x}$ near 2.1 use "Sadhu"verbs.[1/] In prose, the two phenomena are closely correlated. "Devi Chaudhurani" mostly uses"Sadhu" verbs etc., but its style is very nearly colloquial; it has a mean of 2.26, while 'Yogayog' in perfect 'chalita' style shows $\bar{x}$ nearly 2.20. But this only means that the colloquial form in the narrow sense, covering only verbs and pronouns, cannot influence the $\bar{x}$ greatly. Usually, however, colloquial verbs etc. are associated with a really colloquial language, and 'sadhu'verbs etc. with a really elevated language, and the choice of one of the two styles does influence the $\bar{x}$ greatly.

3.3.8. Historical novels tend to have higher values of $\bar{x}$, between 2.4 and 2.6, while the other novels generally have values below 2.4. Journalistic writing, represented by "Dristipat", may also have a somewhat higher average.

1/ A poem with a high $\bar{x}$ is generally on a serious theme, but the converse is not true, that is, a low value of $\bar{x}$ does not preclude the poem from being serious; for example, "Ami" with $\bar{x}$ = 2.18 is highly philosophical.

3.3.9. It will be seen in the following chapter that conversational passages have shorter words, on the average, than words in non-conversational matter : the difference varies from 0.1 or 0.2 to 0.5 or 0.6 syllables per word, depending on whether the sadhu or the chalita style is used within conversations and outside. This means that the overall $\bar{x}$ tends to be lower for a work with a higher proportion of conversational matter. This is the reason why, for example, "Birbaler Halkhata", a collection of essays, has a much higher $\bar{x}$ than "Char-Yari Katha", a fiction; both were written in the colloquial style by a champion of the colloquial style, Pramatha Chaudhury. Essays can hardly have anything conversational, while fiction generally has. So essays tend to have higher averages than fiction (novels and short stories). Within fiction, again, the value of $\bar{x}$ tends to be lower when the proportion of conversational matter is high. The average for "Kabuliwalla" is so high mainly because it includes very little of conversation. (More of this in Chapter 5.)

3.3.10. Some works written as speeches or thoughts of the leading character(s) tend to show low values of $\bar{x}$. Examples of such cases are "Ghare Baire", "Char-Yari Katha", "Deshe Videshe" and "Chacha Kahini". One might say that in these cases, the major explanation of the low value of $\bar{x}$ would be the use of the colloquial style throughout. But the decision to use the colloquial style has to be taken twice, in general, once for conversations and again for the remaining narrative. Since conversation usually accounts for well below 50% of all words and

quite often only the conversations are written in the colloquial style, the overall $\bar{x}$ is really low only if all words are in the colloquial style. This is ensured if the work is written as speeches or thoughts of one or more leading characters.

3.3.11. From the data on word-length in conversational passages in the most colloquial style [vide Chapter 4, Section 4.4 and Chapter 5, Section 5.5] and the poems for children ["Virpurus" and "Khelabhola" in Table 3.3] it is felt that any non-trivial writing in Bengali cannot have average word-length appreciably below 2 syllables per word.

3.4.1. Within author differences : It has been seen how the two leading writers of Bengali prose — Bamkimchandra Chatterjee and Rabindranath Tagore — show declining trends in $\bar{x}$ in their different novels. Some other authors also show appreciable and significant differences between their different works; out of these, the figures for Bibhutibhusan are specially noteworthy as all the three works come under novels. This seems to be a major finding of the present study, although one in a negative sense.

3.4.2. Fractile graphical analysis is not particularly suited to the analysis of word-length in syllables, the distributions being rather too discrete. Nevertheless, it would be useful to convert the word-length distributions of Tables 3.2 and 3.4 into the fractile form and examine the fractile graphs thus obtained. This is done in Table 3.5 and Figs.3.1 (a)-(b). Table 3.5 relates to some representative works selected from those covered in Table 3.2, and shows the decile

group averages of word-length, separately for the entire probability or pooled sample of words, as well as for the two halfsamples comprising this sample. "Shakuntala" (1854) by Vidyasagar represents the high level (Sanskritised style) at which Bengali fiction started, and "Sheser Kavita" (1929) if not "Ghare Baire" (1916), the level (colloquial style) reached at the end of the Tagore period. Figs. 3.1(a)-(b) are based on Table 3.5 and bring out the within author variation by means of illustrative fractile graphs.

3.4.3. Fig. 3.1(a) relates to two works by Bankimchandra, "Durgeshnandini" (1865), his first novel, and "Anandamath" (1882), a late one, both coming under historical romance. The fractile graphs show clearly significant separation. "Devi Chaudhurani" (1884) would have presented even lower fractile graphs, but one might argue whether it is strictly comparable with "Durgeshnandini" or "Anandamath". Fig.3.1(b) relates to three novels by Tagore. The separation between any two of the works is significant beyond doubt.1/ "Gora" employs the chaste style in non-conversational matter, and the colloquial style in conversations. The other two are entirely in the colloquial style. But apart from this, "Gora" and "Ghare Baire" are not very different in subject-matter.

3.4.4. Broadly speaking, fractile graphs corroborate the conclusions reached on the basis of $\bar{x}$; the work with a higher value of $\bar{x}$ generally shows a higher set of fractile graphs.

1/ "Rajarsi" would present graphs even higher than those of "Gora", but it falls under the historical category and has therefore been omitted.

Table 3.5 Decile group averages of word-length in syllables, based on probability/pooled* samples of words, for selected works in Bengali prose, separately by half-samples and combined.

work	type of sample	half sample	no.of sample words	average length by fractile groups (%) 0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	0-100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1. Shakuntala	prob.	1	351	1.000	1.661	2.000	2.000	2.213	3.000	3.000	3.000	3.908	4.739	2.652
		2	345	1.014	2.000	2.000	2.000	2.536	3.000	3.000	3.087	4.000	4.928	2.756
		comb.	696	1.000	1.836	2.000	2.000	2.372	3.000	3.000	3.000	3.996	4.831	2.704
2. Durgeshnandini	pooled*	1	1190	1.000	1.697	2.000	2.000	2.000	2.949	3.000	3.000	3.672	4.561	2.588
		2	1169	1.000	1.494	2.000	2.000	2.000	2.825	3.000	3.000	3.618	4.945	2.588
		comb.	2359	1.000	1.597	2.000	2.000	2.000	2.888	3.000	3.000	3.645	4.752	2.588
3. Anandamath	pooled*	1	983	1.000	1.474	2.000	2.000	2.000	2.232	3.000	3.000	3.312	4.539	2.456
		2	927	1.000	1.220	2.000	2.000	2.000	2.196	3.000	3.000	3.316	4.518	2.425
		comb.	1910	1.000	1.351	2.000	2.000	2.000	2.215	3.000	3.000	3.314	4.529	2.441
4. Gora	pooled*	1	1365	1.000	1.484	2.000	2.000	2.000	2.000	2.809	3.000	3.000	4.230	2.352
		2	1348	1.000	1.368	2.000	2.000	2.000	2.000	2.546	3.000	3.046	4.324	2.328
		comb.	2713	1.000	1.426	2.000	2.000	2.000	2.000	2.678	3.000	3.000	4.299	2.341
5. Ghare Baire	prob.	1	984	1.000	1.000	1.835	2.000	2.000	2.000	2.000	2.134	3.000	3.699	2.067
		2	917	1.000	1.037	2.000	2.000	2.000	2.000	2.000	2.290	3.000	3.873	2.120
		comb.	1901	1.000	1.000	1.933	2.000	2.000	2.000	2.000	2.209	3.000	3.782	2.093
6. Sheser Kavita	pooled*	1	1013	1.000	1.243	2.000	2.000	2.000	2.000	2.000	2.822	3.000	4.054	2.212
		2	1006	1.000	1.052	2.000	2.000	2.000	2.000	2.000	2.654	3.000	4.125	2.183
		comb.	2019	1.000	1.148	2.000	2.000	2.000	2.000	2.000	2.739	3.000	4.092	2.198

* "Pooled" means probability plus systematic.

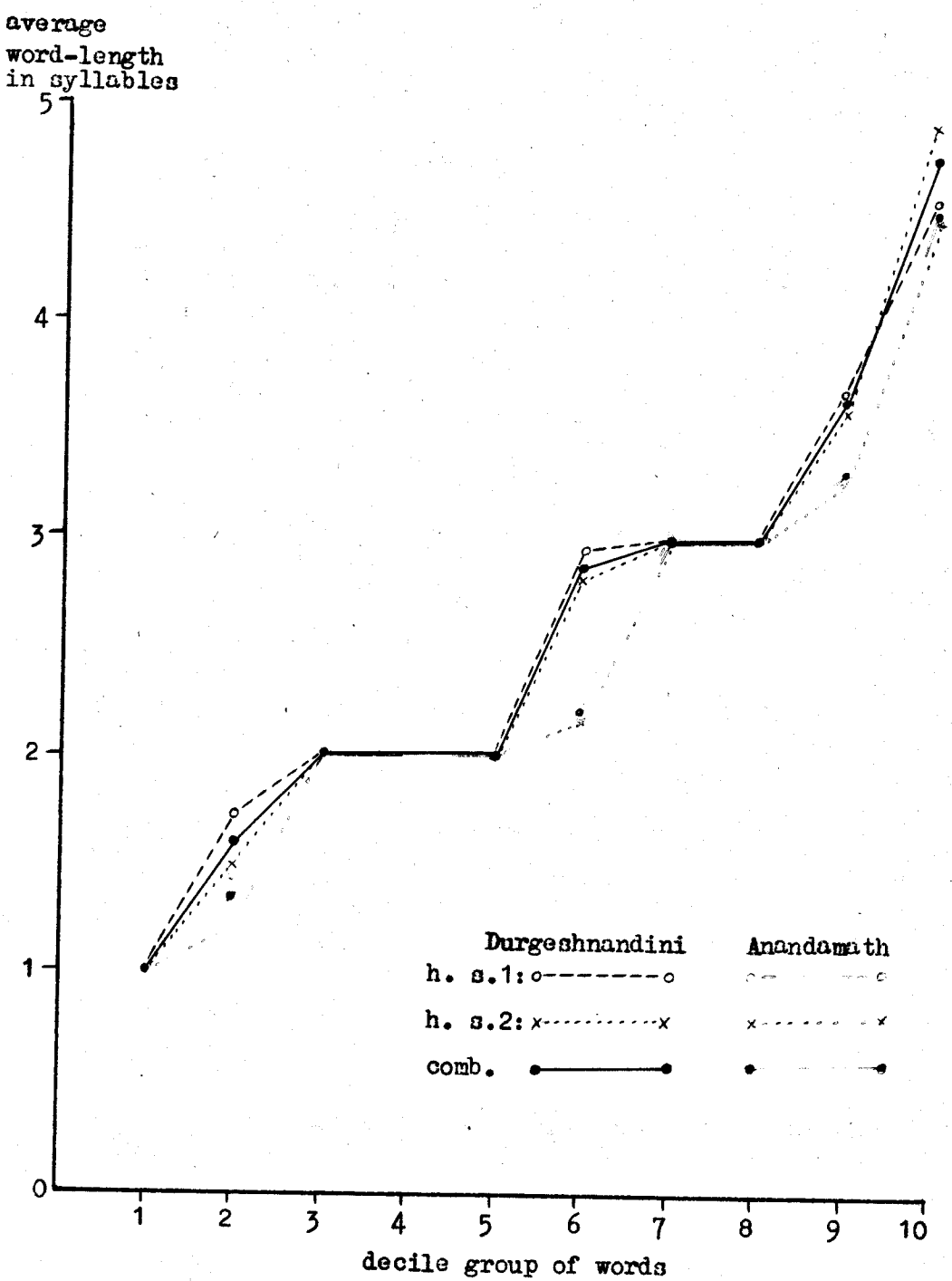


Fig.3.1(a): Fractile graphs for word-length in syllables for two selected works by Bamkimchandra, based on 'pooled',i.e., probability plus systematic, samples of words : "Durgeshnandini" - 2359 sample words and "Anandamath" - 1910 sample words.

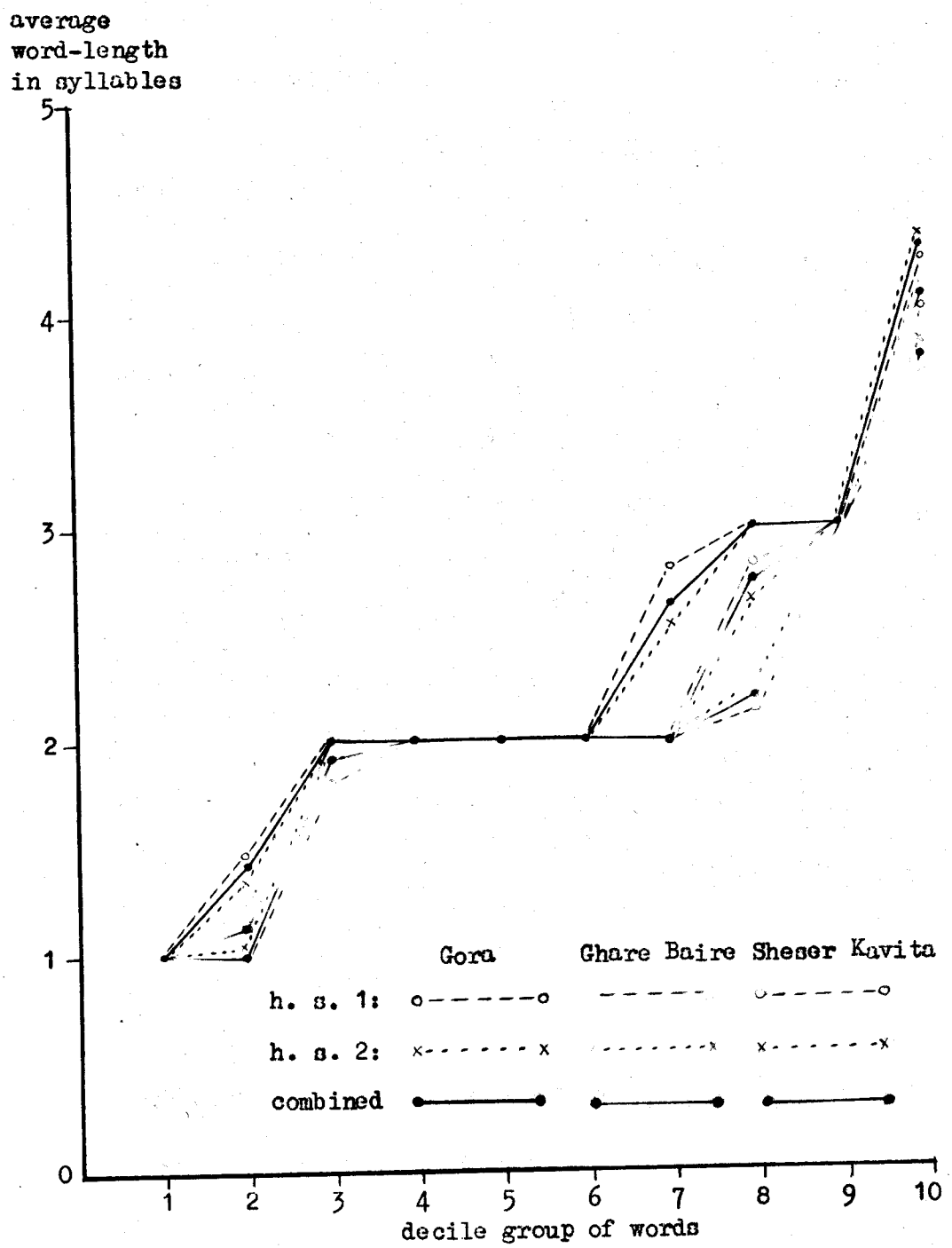


Fig. 3.1(b): Fractile graphs for word-length in syllables, for three selected works by Rabindranath Tagore, based on samples of words : "Gora" -'pooled' sample (2713 words), "Ghare Baire" - probability sample (1901 words) and "Sheser Kavita" - 'pooled' sample (2019 words).

3.4.5. Statistical investigations of style have generally aimed at defining suitable statistical indicators of style, like distributions of word-length or sentence-length or diversity of vocabulary, and at showing that these indicators have similar values for different works by any given author, while the values for different authors tend to occupy different non-overlapping ranges. If this were true, in general, one would indeed possess dependable statistical indices, which could form the basis of objective answers to problems of disputed authorship, say. Some statistical investigations on western languages seem to give an optimistic picture. [Vide Yule, 1938, 1944; Fucks, 1952; Mendenhall, 1887, 1901 (reviewed by Williams, 1956)].[1/]

3.4.6. But so far as distributions or averages of word-length are concerned, the situation seems to be very different in Bengali prose or poetry.[2/] It is true that the present study emphasises the formative period of modern Bengali prose, when the literary language was changing fast, so much so that in 100 years, the language lived through processes undergone by the English language (say) in course of several centuries. But from common knowledge it may be said that even today, when Bengali prose may be said to have attained maturity, the same Bengali author sometimes writes two works in very different languages. [Compare, for instance, "Triyama", a novel and "Bharat Premkatha", a collection of short stories, both by Subodh Ghosh.] Some authors may

1/ See also Chapter 1, Section 1.2, especially the concluding paras.

2/ A similar negative conclusion will be reached for sentence-length in Bengali prose (vide Chapter 8).

have a stable style, but this cannot be assumed for all, unless one confines one's attention to a very narrow field of literature. Within author differences may be appreciable in Bengali prose even today; and poetry is, for reasons given, even more flexible. Word-length data cannot indicate the authorship of a Bengali work with any degree of certainty.

3.4.7. It may well be that within author variation is much smaller in English and other languages, where the literary language does not change so rapidly during the life-time of a single author. But still it seems that previous researchers have tried to draw far-reaching generalisations on insufficient grounds.[1] Indeed, studies now classical show the variation of Plato's and Shakespeare's styles with age. The style of an author, as shown by word-length etc, can certainly change with his age and also most certainly depends on the topic on which he writes, i.e., the particular field of literature. Some cautious statements like this are sometimes made. But if this were true, the usefulness of word-length etc. for showing the individual style of the author is considerably less than if there were no such variation within the author.

3.4.8. But even this restrained form of conclusion does not seem to be justified for Bengali prose during the last hundred years. Compare, for example, "Pather Panchali", "Aparajita" and "Devayan" by Bibhutibhusan, a post-Tagore writer. Among Tagore's works, one may

1/ Inductive inference has always this risk that the conclusion may change as and when new evidence is available.

compare "Ghare Baire" (1916) and "Chaturanga" (1916), and among Bamkimchandra's, one may compare "Visavriksha" (1873) with "Krishnakanter Will" (1878) or "Anandamath" (1882) with "Devi Chaudhurani" (1884). In each case, the two works of the pair are fairly similar as regards subject-matter and were also written at not very distant dates. Still, there are considerable and significant differences in the word-length distributions.

3.4.9. The difference between "Pather Panchali", "Aparajita" and "Devayan" will be partly explained in the following chapter by the considerably higher proportion of conversational matter in "Devayan". One may also argue that the subject-matter is somewhat different in "Devayan". The difference between "Ghare Baire" and "Chaturanga" can, of course, be ascribed to the sudden change-over from the chaste to the colloquial style. "Anandamath" and "Devi Chaudhurani" differ in elevation or seriousness. In this way, one can go on explaining the observed differences, one after another, pointing to some difference in the subject-matter or at least in the mood or elevation, if nothing else. But then the statement that an author has a characteristic type of word-length distribution for a given type of literature at a given age has very little content. And in fact the choice of the level of average word-length itself contributes greatly to the mood or elevation. This is the truth about the difference between "Visavriksha" and Krishnakanter Will" which are so very similar in subject-matter and were written at an interval of only five years.

3.4.10. Not only does the word-length distribution depend on the type of literature, but it can also change with the age of the author; also during the formative period of Bengali prose, it sometime varied erratically between similar works written by the same author at about the same period of time.

3.5.1. Word-length distributions — other aspects : We may first consider the position in Bengali prose. The percentage of monosyllables varies from about 11 ("Shakuntala") to 23 ("Char-Yari Katha"). The percentage of bisyllabled words ranges between 31 ("Sitar Vanavas") and 57 ("Ghare Baire"). Similarly, p_3 lies between 34% and 14%, and p_4 between 14% and 4%.

3.5.2. For the two works by Vidyasagar in the chaste style, having the highest values of average word-length, the distribution is broadly as follows:

length in syllables :	1	2	3	4	5	6 or more	total
percentage of words :	13	33	34	12	5	3	100

The frequencies of two-syllabled and three syllabled words are nearly equal. On the other hand, for the latest works entirely in colloquial style, the percentage distribution is roughly as follows:

length in syllables :	1	2	3	4	5	6 or more	total
percentage of words :	18	53	20	7	2	<1	100

These percentages are averages over "Devayan", "Dristipat", "Janantik", "Deshe Videshe", "Chacha Kahini", "Vishwavidyalay" and "Laboratory". These works, it may be noted, include belles letters and essays. The

lowest values of average word-length are shown by "Char-Yari Katha" and "Ghare Baire", both entirely in the colloquial style, but not included in the above mentioned set of seven recent works. The average distribution for these two works is shown below:

length of syllables :	1	2	3	4	5	6 or more	total
percentage of words :	22	57	15	5	1	<1	100

Clearly, in modern Bengali prose the mode of the word-length distribution is at 2 syllables, which reminds us of the "law of bimorism" (Chatterjee, 1945, p.38).

3.5.3. The standard deviations (s_x) for prose works range between 0.83 in "Ghare Baire" and 1.25 in "Samya". The values of s_x rise steadily, being highly correlated, with the corresponding values of $\bar{x}$.[1/] For works with $\bar{x}$ near 2.1, the value of s_x is usually around 0.85, so that coefficient of variation is 40%; for works with $\bar{x}$ near 2.7, the value of s_x tends to be nearly 1.2, which means C.V. is nearly 45%.

3.5.4. As regards poetry, the percentages p_r are seen to cross the limits quoted in para 3.5.1. Thus, p_1 is only 3.4% in "Varshamangal" where $\bar{x}$ is highest (3.35); p_2 is about 65 or 70% in the three poems with the lowest values of $\bar{x}$, viz., "Krishnakali", "Virpurus" and "Khelabhola". Similarly, p_3 is 38% in "Varshamangal" and 5 to 8% in the three other poems just named; and so on. One explanation is that "Varsamangal" and

---

On the whole, the distributions of x seem to shift systematically with changes in $\bar{x}$, forming, roughly, a uniparameter family of distributions (vide, however, Chapter 6).

two other poems have higher values of $\bar{x}$ than any prose work covered in Tables 3.1 and 3.2. But it also appears that poems in playful, colloquial style; with very low values of $\bar{x}$ have larger p_2 and smaller p_3 than prose works in colloquial style (like "Ghare Baire") having equally low values of $\bar{x}$. Possibly, metrical considerations favour bisyllabic words in preference to trisyllables.

3.5.5. The close correlation between $\bar{x}$ and s_x is observed in poetry also. The observed range of $\bar{x}$ is wider for poetry than in prose, viz., from 1.95 to 3.35, and the corresponding limits for s_x are even wider, viz., 0.65 to 1.57. The C.V. tends to increase with $\bar{x}$ from about 35 to 45%. The poem "Ora Kaj Kare" is exceptional, having a moderate $\bar{x}$ with a relatively high s_x, so that C.V. is 57%.

3.5.6. Appendix 2 shows the usefulness of concentration curves based on word-length distributions.

Chapter 4: Words within and outside conversations

4.1.1. Introduction : When the study on word-length had advanced to some extent, it became apparent that a classification of words into "conversational" and "others" could lead to a better understanding of the between works variation in word-length. As an illustration, we may refer to the works by Bibhutibhusan Bandyopadhyay viz., "Pather Panchali", "Aparajita", and "Devayan". The averages of word-length are given below for all three works, separately for conversational words ($\bar{x}_c$), for other words ($\bar{x}_o$) and for all words ($\bar{x}$). The percentages of all words falling under "conversation" (p_c) are also given.[1/] Clearly, $\bar{x} = \frac{p_c \bar{x}_c + (100 - p_c) \bar{x}_o}{100}$.

work	no. of sample words	p_c	$\bar{x}_c$	$\bar{x}_o$	$\bar{x}$
(1)	(2)	(3)	(4)	(5)	(6)
Pather Panchali	2552	26.21	1.912	2.396	2.269
Aparajita	1894	26.35	2.000	2.371	2.274
Devayan	3176	55.20	2.093	2.193	2.138

Average word-length is more or less the same in both classes of words, and also for all words, so far as the first two works are concerned; but "Devayan" seems to have longer words in its conversational matter, and shorter words in its non-conversational passages than the other two works.

1/ These estimates are reproduced from Tables 4.1 and 4.2 where estimates for all works are given by type of sample and by subsamples.

4.1.2. Both "Pather Panchali" and "Aparajita" have p_c of the order of 25%, but for "Devayan" the percentage is much higher, about 55. In a sense, this is responsible for the considerable and significant differences in the overall averages ($\bar{x}$) between either "Pather Panchali" or "Aparajita", on the one hand, and "Devayan", on the other. If the first two works had the same value of p_c as 'Devayan,' while $\bar{x}_c$ and $\bar{x}_o$ were as actually observed, the overall averages $\bar{x}$ would have been :

Pather Panchali : $0.552 \times 1.912 + 0.448 \times 2.396 = 2.129$

Aparajita : $0.552 \times 2.000 + 0.448 \times 2.371 = 2.166$

These averages are closer to the average for 'Devayan' (2.138) than the actual averages for "Pather Panchali" (2.269) and "Aparajita" (2.274).[1/]

4.1.3. Evidently, the classification of words into "conversational" and "others" gives a more detailed analysis of word-length. Conceivably, two works may have similar values of $\bar{x}_c$ and $\bar{x}_o$, but the $\bar{x}$-values may be quite different due to unequal weightages of conversations. The classification is, of course, applicable to fiction, but in these cases, it has much wider significance than brought out above.

4.1.4. First of all, the percentage of conversational matter (p_c) may itself be regarded as an indicator of style. And second, instead of measuring word-length in fiction by one overall distribution or average, it is better to use two distribution and two averages ($\bar{x}_c$ and $\bar{x}_o$), one

---

1/ If, however, "Devayan" had the same value of p_c as "Pather Panchali" or "Aparajita", it would have a value of $\bar{x}$ only about 2.17, and the gap would remain practically the same.

for conversational words, and the second for the remaining words. This is because a writer of fiction, in Bengali at least, has to make two distinct choices regarding the level of language to be employed, one for the language to be used in conversations, and the other for the remaining narrative. The problem may not be very important today, but it was very important during the formative period of Bengali prose. The statement below shows the works employing (i) the chaste style ("Sadhu Bhasa") in both situations; (ii) the chaste style outside conversations and the colloquial style ("Chalita Bhasa") for conversations; and (iii) the "chalita" style throughout the text.

level of language used		works
in conversations	outside conversations	
chaste	chaste	Shakuntala, Sitar Vanavasa, Durgeshnandini, Kapalkundala, Visavriksha, Krishnakanter Will, Anandamath, Devi Choudhurani, Bouthakuranir Hat, Rajsinha, Rajarsi, Chokher Bali, Chaturanga, Kabuliwalla, Kshudhita Pasan.
colloquial	chaste	Gora, Pallisamaj, Pather Dabi, Pather Panchali, Aparajita.
colloquial	colloquial	Ghare Baire, Sheser Kavita, Yogayog, Laboratory, Char-Yari Katha, Birbaler Halkhata, Devayan, Dristipat, Janantik, Chacha-Kahini, Deshe Videshe.

N.B.: (1) In several works by Bamkimchandra, particularly "Devi Choudhurani", colloquial forms are sometimes used in conversations.
(2) "Chaturanga" is written in a nearly colloquial style, although the verbs and pronouns have the chaste form.
(3) "Kabuliwalla" occasionally uses the colloquial style in conversational matter.

4.1.5. The present chapter introduces this simple division of words in the text and studies the word-length distributions for prose fiction given in Chapter 3 separately for the two classes of words. The emphasis is on the two averages.[1] The percentage (p_c) of conversational words is also studied. This throws much light on the field of study. Some of the main findings were mentioned in advance in Chapter 3. But too much should not be expected, either. Many of the between works differences remain unexplained; that is to say, the averages $\bar{x}_c$ and $\bar{x}_o$ also show considerable variation between works and even within authors. Also, the sampling errors of the estimates of p_c are often very large (for reasons see below); the sampling errors of $\bar{x}_c$ and $\bar{x}_o$ are also naturally larger than those of the corresponding estimates $\bar{x}$. So not many definite conclusions can be reached about the variation in these quantities.

4.1.6. It will, nevertheless, be evident beyond doubt that in most of the Bengali works coming under "fiction", $\bar{x}_o$ is significantly and appreciably larger than $\bar{x}_c$. The difference varies from about 0.1 (syllables) for works like "Ghare Baire" which employ the colloquial style throughout to about 0.4 or 0.5 (syllables) for works falling in the two other groups shown in the tabular statement in para 4.1.4.

4.1.7. In Chapter 5 we study the randomness of the series of word-lengths by means of autocorrelation coefficients and by other techniques. It will be seen that in Bengali prose the series of word-lengths is very nearly random; the autocorrelation coefficients of the first

1/ The difference between the two averages is evidently of great interest.

few orders (r_1, r_2, etc.) are of the order of 0.05 if not nearer zero. But one non-random feature of the series is being demonstrated in the present chapter. It does not need any statistical test to prove that either conversational words or non-conversational words, when they occur, occur in long runs, that is, the two classes of words do not alternate so frequently as predicted by the theory of runs of two kinds of elements in a random series (Wald and Wolfowitz, 1940)
Now if $\bar{x}_o$ is larger than $\bar{x}_c$, this would lead to alternate patches (long runs) with longish nonconversational words and shortish conversational words, and shifts in the average from one patch to the next. Such shifts would be conspicuous in works with larger values of $\bar{x}_o - \bar{x}_c$, e.g., "Sitar Vanavas", but the contrast is not pronounced when $\bar{x}_o$ is only slightly larger than $\bar{x}_c$, as in "Ghare Baire". This aspect will be taken up again in Chapter 5.

4.2.1. The material : What now is "conversational matter" ? The choice of a definition was not very obvious. If the emphasis is on matter written in the conversational style with shortish words, like colloquial forms of verbs and pronouns, one could include letters, at least short ones, soliloquies, words addressed to gods or to absent persons, words spoken in dreams etc., within the category of conversation. But these boderline cases were all left out and a narrow definition adopted here for the sake of convenience : "Conversational matter" was taken to include only words actually uttered in conversation with persons present. It is felt that a wider definition would not lead to

very different results. But it would probably have been better to form a third category with these borderline cases, making the "others" category purer.[1/]

4.2.2. Some of the works, viz., "Chaturanga", "Ghare Baire", "Char-Yari Katha", "Dristipat", "Chacha Kahini" and "Deshe Videshe" are written as thoughts or speeches of the author or of some leading characters, who are mentioned in the first person. This lends a conversational character to even the non-conversational matter in these works.

4.2.3. Table 4.1 presents the estimates p_c for different works in Bengali prose studied in Chapter 3, excluding the essays and collections of essays, but including the belles letters which contain conversational matter. The estimates are given by type of sample and also by subsamples. Table 4.2 shows the corresponding averages $\bar{x}_c$ and $\bar{x}_o$, again by type of sample and by subsamples. Table 4.3 presents the word-length distributions by works and by type of sample, separately for conversational words and other words; sub-samplewise details are not shown here for lack of space, but where two types of samples are available, the purpose may be served by them. The sample sizes in terms of words have been shown in all these tables as far as possible.

4.2.4. No standard error has been calculated for these estimates. The inferences will be based on the subsamplewise estimates.

1/ This remark applies, in particular, to "Visavriksha" where this third category is fairly important.

Table 4.1: Percentage of 'conversational' words[1/] estimated for selected works in Bengali prose, separately by type of sample and by subsamples.

author	work	type of sample	total no. of words by subsamples					percentage of conversational words				
			ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Vidyasagar	Shakuntala	prob.	170	181	164	181	696	57.65	53.59	35.37	59.12	51.72
	Sitar Vanavas	prob.	193	191	187	179	750	35.75	62.30	47.59	48.04	48.40
Bamkimchandra	Durgeshnandini	prob.	148	147	129	153	577	12.84	34.01	27.91	25.49	24.96
		syst.	487	408	424	463	1782	23.20	32.11	33.49	28.08	28.96
		pooled	635	555	553	616	2359	20.79	32.61	32.19	27.44	27.98
	Kapalkundala	syst.	104	135	130	124	493	23.08	37.04	25.38	33.87	30.22
	Visavriksha	prob.	151	146	147	167	611	28.48	5.48	17.01	8.38	14.73
		syst.	505	431	448	468	1852	17.03	17.87	16.96	18.16	17.49
		pooled	656	577	595	635	2463	19.67	14.73	16.97	15.59	16.81
	Krishnakanter Will	prob.	442	427	456	452	1777	28.05	29.98	31.14	25.88	28.76
		syst.	192	194	171	192	749	29.17	27.84	40.94	44.79	35.51
		pooled	634	621	627	644	2526	28.39	29.31	33.81	31.52	30.76
	Anandamath	prob.	286	294	270	259	1109	36.36	45.24	46.67	38.61	41.75
		syst.	197	206	186	212	801	29.44	23.30	30.11	39.62	30.71
		pooled	483	500	456	471	1910	33.54	36.20	39.91	39.07	37.12
	Devi Choudhurani	prob.	283	283	317	291	1174	33.92	50.88	38.49	36.77	39.95
		syst.	203	195	234	201	833	36.45	31.80	49.57	52.74	42.98
		pooled	486	478	551	492	2007	34.98	43.10	43.19	43.29	41.21
	Rajsinha	prob.	334	368	359	362	1423	41.02	40.76	30.36	37.02	37.25
		syst.	130	123	131	123	507	30.00	39.84	40.46	34.96	36.29
		pooled	464	491	490	485	1930	37.93	40.53	33.06	36.50	36.99

1/ 'Conversational' words are those actually uttered in conversation with persons present.

4.1: (Contd.)

author	work	type of sample	total no. of words by subsamples					percentage of conversational words				
			ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Rabindranath	Bouthakuranir Hat	prob.	408	405	406	373	1592	38.97	35.80	51.48	43.16	42.34
		syst.	211	212	218	186	827	25.59	38.68	56.88	36.56	39.66
		pooled	619	617	624	559	2419	34.41	36.79	53.36	40.97	41.42
	Rajarsi	prob.	386	420	421	405	1632	34.97	15.00	26.84	25.19	25.31
		syst.	172	183	176	158	689	31.40	32.24	21.02	8.86	23.80
		pooled	558	603	597	563	2321	33.87	20.23	25.13	20.60	24.86
	Chokher Bali	syst.	337	308	322	351	1318	28.18	31.49	24.53	31.05	28.83
	Gora	prob.	213	221	216	239	889	37.09	47.96	33.33	40.17	39.71
		syst.	492	439	473	420	1824	43.49	34.40	34.67	36.43	37.39
		pooled	705	660	689	659	2713	41.56	38.94	34.25	37.78	38.15
	Chaturanga	prob.	341	356	386	375	1458	17.01	25.84	25.13	21.87	22.57
		syst.	216	219	215	204	854	20.83	21.46	25.12	23.04	22.60
		pooled	557	575	601	579	2312	18.49	24.17	25.12	22.28	22.58
	Ghare Baire	prob.	509	475	417	500	1901	30.05	21.47	25.18	34.00	27.88
	Sheser Kavita	prob.	188	178	185	184	735	40.42	38.76	42.70	61.96	45.99
		syst.	307	340	311	326	1284	48.86	52.94	59.49	56.75	54.52
		pooled	495	518	496	510	2019	45.66	48.07	53.22	58.63	51.41
	Yogayog	syst.	293	307	294	293	1187	29.69	26.71	34.35	32.76	30.83
Pramatha Choudhury	Char-Yari Katha	prob.	224	213	218	217	872	29.02	34.74	37.16	35.94	34.17

Table 4.1 : (Contd.)

author	work	type of sample	total no. of words by subsamples					percentage of conversational words				
			ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Saratchandra	Pallisamaj	prob.	218	228	215	229	890	45.87	32.89	39.07	49.78	41.91
	Pather Dabi	prob.	220	197	202	196	815	34.09	50.76	46.04	43.37	43.31
Bibhutibhusan	Pather Panchali	prob.	222	239	233	228	922	35.59	36.82	36.05	21.05	32.43
		syst.	422	436	397	375	1630	25.12	23.17	22.17	20.00	22.70
		pooled	644	675	630	603	2552	28.73	28.00	27.30	20.40	26.21
	Aparajita	syst.	511	484	452	447	1894	24.07	37.40	29.20	14.09	26.35
	Devayan	prob.	233	236	239	223	931	62.66	62.29	71.97	42.15	60.04
		syst.	562	559	574	550	2245	53.20	46.87	54.53	58.18	53.18
		pooled	795	795	813	773	3176	55.97	51.45	59.65	53.56	55.20
Jajabar	Dristipat.	prob.	188	194	199	191	772	12.77	5.67	22.11	10.99	12.95
		syst.	402	409	386	394	1591	4.73	11.98	19.69	9.39	11.38
		pooled	590	603	585	585	2363	7.29	9.95	20.51	9.91	11.89
	Janantik	prob.	163	183	171	173	690	26.99	37.16	12.87	24.28	25.51
Muztaba Ali	Chacha-Kahini	prob.	189	205	195	185	778	20.11	19.02	8.21	24.34	17.87
	Deshe Videshe	prob.	208	182	193	208	791	46.15	37.91	37.31	21.15	35.52
Rabindranath	Kabuliwalla	syst.	198	195	191	195	779	2.02	19.49	6.81	13.85	10.53
	Kshudhita Pasan	syst.	307	288	297	300	1192	3.26	2.78	4.04	7.00	4.28
	Laboratory	syst.	318	283	321	306	1228	67.30	64.66	70.09	72.88	68.81

Table 4.2: Average length of 'conversational words' and 'other words' estimated for selected works in Bengali prose, separately by type of sample and by subsamples.*

author	work	type of sample	average length in syllables: conversational words by subsamples					other words by subsamples					all words
			ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.	(comb.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Vidyasagar	Shakuntala	prob.	2.735 (98)	2.330 (97)	2.638 (58)	2.542 (107)	2.553 (360)	2.986 (72)	2.643 (84)	2.991 (106)	2.824 (74)	2.866 (336)	2.704 (696)
	Sitar Vanavas	prob.	2.493 (69)	2.613 (119)	2.551 (89)	2.663 (86)	2.587 (363)	2.710 (124)	2.750 (72)	2.776 (98)	2.968 (93)	2.796 (387)	2.695 (750)
Bamkim-chandra	Durgesh-nandini	prob.	2.632 (19)	2.200 (50)	2.750 (36)	2.026 (39)	2.347 (144)	2.690 (129)	2.598 (97)	2.839 (93)	2.518 (114)	2.656 (433)	2.579 (577)
		syst.	2.460 (113)	2.359 (131)	2.493 (142)	2.385 (130)	2.424 (516)	2.634 (374)	2.704 (277)	2.642 (282)	2.664 (333)	2.659 (1266)	2.591 (1782)
		pooled	2.485 (132)	2.315 (181)	2.545 (178)	2.302 (169)	2.407 (660)	2.648 (503)	2.677 (374)	2.691 (375)	2.627 (447)	2.658 (1699)	2.588 (2359)
	Kapalkundala	syst.	2.333 (24)	2.140 (50)	2.424 (33)	2.071 (42)	2.208 (149)	2.925 (80)	2.882 (85)	2.742 (97)	2.805 (82)	2.834 (344)	2.645 (493)
	Visavriksha	prob.	2.140 (43)	1.750 (8)	1.960 (25)	2.214 (14)	2.067 (90)	2.509 (108)	2.580 (138)	2.648 (122)	2.438 (153)	2.539 (521)	2.470 (611)
		syst.	2.209 (86)	2.299 (77)	2.184 (76)	2.388 (85)	2.272 (324)	2.508 (419)	2.506 (354)	2.564 (372)	2.400 (383)	2.494 (1528)	2.455 (1852)
		pooled	2.186 (129)	2.247 (85)	2.129 (101)	2.363 (99)	2.227 (414)	2.508 (527)	2.527 (492)	2.585 (494)	2.411 (536)	2.505 (2049)	2.459 (2463)

Table 4.2 : (Contd.)

author	work	type of sample	average length in syllables: conv. words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	other words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	all words (comb.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Krishna-kanter Will	prob.	2.194 (124)	2.062 (128)	2.092 (142)	2.026 (117)	2.094 (511)	2.384 (318)	2.150 (299)	2.490 (314)	2.502 (335)	2.439 (1266)	2.340 (1777)
		syst.	1.964 (56)	2.037 (54)	2.229 (70)	2.186 (86)	2.120 (266)	2.537 (136)	2.529 (140)	2.545 (101)	2.424 (106)	2.511 (483)	2.372 (749)
		pooled	2.122 (180)	2.055 (182)	2.137 (212)	2.094 (203)	2.103 (777)	2.430 (454)	2.271 (439)	2.503 (415)	2.483 (441)	2.459 (1749)	2.350 (2526)
	Anandamath	prob.	2.115 (104)	2.241 (133)	2.175 (126)	2.320 (100)	2.212 (463)	2.555 (182)	2.609 (161)	2.778 (144)	2.516 (159)	2.608 (646)	2.443 (1109)
		syst.	2.259 (58)	2.271 (48)	2.429 (56)	1.917 (84)	2.183 (246)	2.612 (139)	2.582 (158)	2.415 (130)	2.461 (128)	2.551 (555)	2.438 (801)
		pooled	2.167 (162)	2.249 (181)	2.253 (182)	2.136 (184)	2.202 (709)	2.580 (321)	2.596 (319)	2.606 (274)	2.491 (287)	2.582 (1201)	2.441 (1910)
	Devi Choudhurani	prob.	2.167 (96)	2.000 (144)	2.066 (122)	2.075 (107)	2.068 (469)	2.449 (187)	2.446 (139)	2.267 (195)	2.554 (184)	2.426 (705)	2.283 (1174)
		syst.	2.216 (74)	2.048 (62)	1.888 (116)	2.057 (106)	2.034 (358)	2.380 (129)	2.346 (133)	2.288 (118)	2.505 (95)	2.373 (475)	2.227 (833)
		pooled	2.188 (170)	2.014 (206)	1.979 (238)	2.066 (213)	2.053 (827)	2.421 (316)	2.397 (272)	2.275 (313)	2.537 (279)	2.405 (1180)	2.260 (2007)
	Rajsinha	prob.	2.402 (137)	2.273 (150)	2.294 (109)	2.321 (134)	2.323 (530)	2.645 (197)	2.601 (218)	2.496 (250)	2.583 (228)	2.577 (893)	2.482 (1423)
		syst.	2.436 (39)	2.326 (49)	2.189 (53)	2.186 (43)	2.277 (184)	2.715 (91)	2.838 (74)	2.512 (78)	2.700 (80)	2.777 (323)	2.596 (507)
		pooled	2.410 (176)	2.286 (199)	2.260 (162)	2.288 (177)	2.311 (714)	2.667 (288)	2.661 (292)	2.500 (328)	2.613 (308)	2.630 (1216)	2.512 (1930)

Table 4.2 . (Contd.)

author	work	type of sample	average length in syllables: conv. words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	other words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	all words (comb.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Boutha-kuranir Hat	prob.	2.283 (159)	2.110 (145)	2.230 (209)	2.267 (161)	2.226 (674)	2.406 (249)	2.500 (260)	2.599 (197)	2.538 (212)	2.504 (918)	2.386 (1592)
		syst.	2.315 (54)	2.183 (82)	2.064 (124)	2.103 (68)	2.143 (328)	2.560 (157)	2.523 (130)	2.532 (94)	2.703 (118)	2.579 (499)	2.406 (827)
		pooled	2.291 (213)	2.136 (227)	2.168 (333)	2.218 (229)	2.199 (1002)	2.466 (406)	2.508 (390)	2.577 (291)	2.597 (330)	2.530 (1417)	2.393 (2419)
	Rajarsi	prob.	2.170 (135)	1.984 (63)	2.265 (113)	2.157 (102)	2.165 (413)	2.514 (251)	2.496 (357)	2.494 (308)	2.548 (303)	2.512 (1219)	2.424 (1632)
		syst.	2.259 (54)	2.458 (59)	1.973 (37)	1.857 (14)	2.232 (164)	2.542 (118)	2.589 (124)	2.460 (139)	2.576 (144)	2.541 (525)	2.467 (689)
		pooled	2.196 (189)	2.213 (122)	2.193 (150)	2.121 (116)	2.184 (577)	2.523 (369)	2.520 (481)	2.483 (447)	2.557 (447)	2.521 (1744)	2.437 (2321)
	Chokher Bali	syst.	2.147 (95)	2.186 (97)	2.177 (79)	2.128 (109)	2.158 (380)	2.401 (242)	2.583 (211)	2.428 (243)	2.409 (242)	2.451 (938)	2.366 (1318)
	Gora	prob.	2.025 (79)	2.151 (106)	2.278 (72)	2.010 (96)	2.110 (353)	2.448 (134)	2.539 (115)	2.486 (144)	2.441 (143)	2.476 (536)	2.331 (889)
		syst.	2.164 (214)	2.166 (151)	2.116 (164)	1.97 (153)	2.110 (682)	2.511 (278)	2.483 (288)	2.489 (309)	2.461 (267)	2.486 (1142)	2.345 (1824)
		pooled	2.127 (293)	2.160 (257)	2.165 (236)	1.988 (249)	2.110 (1035)	2.492 (412)	2.499 (403)	2.488 (453)	2.454 (410)	2.483 (1678)	2.341 (2713)

Table 4.2 : (Contd.)

author	work	type of sample	average length in syllables: conv. words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	other words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	all words (comb.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Cha tu-ranga	prob.	2.379 (58)	2.141 (92)	2.072 (97)	2.159 (82)	2.167 (329)	2.466 (283)	2.303 (264)	2.315 (289)	2.403 (293)	2.373 (1129)	2.326 (1458)
		syst.	2.444 (45)	2.255 (47)	2.148 (54)	2.298 (47)	2.280 (193)	2.322 (171)	2.273 (172)	2.342 (161)	2.268 (157)	2.301 (661)	2.296 (854)
		pooled	2.407 (103)	2.180 (139)	2.099 (151)	2.210 (129)	2.209 (522)	2.412 (454)	2.291 (436)	2.325 (450)	2.356 (450)	2.346 (1790)	2.315 (2312)
	Ghare Baire	prob.	1.948 (153)	2.147 (102)	2.010 (105)	2.106 (170)	2.049 (530)	2.090 (356)	2.072 (373)	2.186 (312)	2.100 (330)	2.109 (1371)	2.093 (1901)
	Sheser Kavita	prob.	2.000 (76)	2.275 (69)	2.114 (79)	2.018 (114)	2.089 (338)	2.205 (112)	2.367 (109)	2.217 (106)	2.300 (70)	2.270 (397)	2.186 (735)
		syst.	2.240 (150)	2.167 (180)	2.168 (185)	2.049 (185)	2.151 (700)	2.293 (157)	2.131 (160)	2.302 (126)	2.305 (141)	2.267 (584)	2.204 (1284)
		pooled	2.159 (226)	2.197 (249)	2.152 (264)	2.037 (299)	2.131 (1038)	2.256 (269)	2.227 (269)	2.263 (232)	2.303 (211)	2.268 (981)	2.198 (2019)
	Yogayog	syst.	2.000 (87)	1.878 (82)	2.010 (101)	2.094 (96)	2.000 (366)	2.184 (206)	2.231 (225)	2.378 (193)	2.188 (197)	2.244 (821)	2.168 (1187)
Pramatha Choudhury	Char-Yari Katha	prob.	1.862 (65)	1.865 (74)	1.963 (81)	2.064 (78)	1.943 (298)	2.126 (159)	2.094 (139)	2.102 (137)	2.158 (139)	2.120 (574)	2.060 (872)
Sarat-chandra	Palli Samaj	prob.	1.930 (100)	1.840 (75)	1.774 (84)	2.000 (114)	1.898 (373)	2.475 (118)	2.333 (153)	2.580 (131)	2.383 (115)	2.439 (517)	2.212 (890)

Table 4.2 : (Contd.)

author	work	type of sample	average length in syllables: conv. words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	other words by subsamples ss 1	ss 2	ss 3	ss 4	comb.	all words (comb.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Sarat-chandra	Pather Dabi	prob.	2.000 (75)	2.140 (100)	1.978 (93)	1.906 (85)	2.011 (353)	2.441 (145)	2.464 (97)	2.358 (109)	2.306 (111)	2.394 (462)	2.228 (815)
Bibhuti-bhusan	Pather Panchali	prob.	1.962 (79)	1.977 (88)	1.964 (84)	1.688 (48)	1.923 (299)	2.462 (143)	2.503 (151)	2.342 (149)	2.339 (180)	2.408 (623)	2.250 (922)
		syst.	1.915 (106)	1.802 (101)	1.909 (88)	2.013 (75)	1.903 (370)	2.418 (316)	2.334 (335)	2.414 (309)	2.397 (300)	2.390 (1260)	2.279 (1630)
		pooled	1.935 (185)	1.883 (189)	1.936 (172)	1.886 (123)	1.912 (669)	2.432 (459)	2.387 (486)	2.391 (458)	2.375 (480)	2.396 (1883)	2.269 (2552)
	Aparajita	syst.	1.959 (123)	2.006 (181)	2.046 (132)	1.968 (63)	2.000 (499)	2.309 (388)	2.393 (303)	2.400 (320)	2.393 (384)	2.371 (1395)	2.274 (1894)
	Devayan	prob.	2.110 (146)	2.000 (147)	2.012 (172)	2.128 (94)	2.054 (559)	2.276 (87)	2.371 (89)	2.179 (67)	2.140 (129)	2.234 (372)	2.126 (931)
		syst.	2.107 (299)	2.149 (262)	2.051 (313)	2.141 (320)	2.111 (1194)	2.106 (263)	2.192 (297)	2.232 (261)	2.187 (230)	2.179 (1051)	2.142 (2245)
		pooled	2.108 (445)	2.095 (409)	2.037 (485)	2.138 (414)	2.093 (1753)	2.148 (350)	2.233 (386)	2.221 (328)	2.170 (359)	2.193 (1423)	2.138 (3176)

Table 4.2 : (Contd.)

author	work	type of sample	average length in syllables										
			conv. words by subsamples					other words by subsamples					all words (comb.)
			ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Jajabar	Dristipat	prob.	2.208 (24)	2.182 (11)	1.864 (44)	2.143 (21)	2.040 (100)	2.421 (164)	2.426 (183)	2.510 (155)	2.412 (170)	2.440 (672)	2.389 (772)
		syst.	2.053 (19)	2.306 (49)	2.053 (76)	1.919 (37)	2.094 (181)	2.418 (383)	2.442 (360)	2.352 (310)	2.529 (357)	2.438 (1410)	2.398 (1591)
		pooled	2.140 (43)	2.283 (60)	1.984 (120)	2.000 (58)	2.075 (281)	2.419 (547)	2.437 (543)	2.405 (465)	2.491 (527)	2.439 (2082)	2.395 (2363)
	Janantik	prob.	2.204 (44)	2.000 (68)	2.136 (22)	2.095 (42)	2.091 (176)	2.429 (119)	2.156 (115)	2.389 (149)	2.450 (131)	2.362 (514)	2.293 (690)
Muztaba Ali	Chacha-Kahini	prob.	2.105 (38)	2.000 (39)	2.125 (16)	2.109 (46)	2.079 (139)	2.252 (151)	2.289 (166)	2.089 (179)	2.238 (143)	2.213 (639)	2.189 (778)
	Deshe-Videshe	prob.	2.052 (96)	2.217 (69)	2.194 (72)	1.796 (44)	2.085 (281)	2.214 (112)	2.212 (113)	2.231 (121)	2.213 (164)	2.220 (510)	2.172 (791)
Rabindra-nath	Kabuliwala	syst.	1.750 (4)	2.053 (38)	2.308 (13)	2.444 (27)	2.207 (82)	2.454 (194)	2.612 (157)	2.556 (178)	2.536 (168)	2.535 (697)	2.501 (779)
	Kshudhita Pasan	syst.	2.200 (10)	1.625 (8)	2.500 (12)	1.667 (21)	1.961 (51)	2.465 (297)	2.629 (280)	2.505 (285)	2.602 (279)	2.549 (1141)	2.524 (1192)
	Laboratory	syst.	2.089 (214)	1.978 (183)	2.067 (225)	2.022 (223)	2.041 (845)	2.404 (104)	2.340 (100)	2.229 (96)	2.337 (83)	2.329 (383)	2.131 (1228)

* Figures within brackets show the number of sample words on which the average length is based.

Table 4.3 Distribution of conversational words (c) and other words (o) by length in syllables estimated for selected works in Bengali prose, separately by type of sample.

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Vidyasagar	Shakuntala	prob.	c	360	14.17	40.82	28.06	11.67	3.33	1.67	0.28		
			o	336	8.93	27.98	39.88	15.77	5.65	1.79			
	Sitar Vanavas	prob.	c	363	17.91	33.05	30.58	11.85	4.68	1.65	-	0.28	
			o	387	12.40	29.72	37.73	11.11	6.20	1.55	1.03	-	0.26(10)
Bamkim-chandra	Durgeshnandini	prob.	c	144	20.83	40.29	25.69	9.72	3.47				
			o	433	13.63	33.95	35.56	10.16	4.39	1.62	0.46	0.23	
		syst.	c	516	16.28	42.83	26.55	11.24	2.71	0.39			
			o	1266	12.48	35.46	34.52	11.77	3.87	0.95	0.79	0.16	
		pooled	c	660	17.27	42.28	26.36	10.91	2.88	0.30			
			o	1699	12.77	35.07	34.79	11.36	4.00	1.12	0.71	0.18	
	Kapalkundala	syst.	c	149	22.82	44.97	24.16	5.37	2.01	0.67			
			o	344	12.79	32.27	29.65	13.37	8.72	2.62	0.29	0.29	
	Visavriksha	prob.	c	90	22.22	55.56	17.78	2.22	2.22				
			o	521	15.16	38.01	30.71	11.52	3.45	0.77	0.38		
		syst.	c	324	17.90	47.22	25.00	9.57	0.31				
			o	1528	15.32	42.33	27.61	9.23	4.06	0.92	0.39	0.07	0.07
		pooled	c	414	18.84	49.03	23.43	7.97	0.73				
			o	2049	15.28	41.24	28.40	9.81	3.90	0.88	0.39	0.05	0.05

Table 4.3 Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bamkim-chandra	Krishnakanter Will	prob.	c	511	23.48	49.91	21.53	3.91	1.17				
			o	1266	17.46	39.89	29.23	9.32	3.40	0.39	0.24	-	0.07
		syst.	c	266	27.44	41.36	25.56	3.38	1.88	0.38			
			o	483	16.98	37.68	30.23	9.73	4.14	0.41	0.62	0.21	
		pooled	c	777	24.84	46.97	22.91	3.73	1.42	0.13			
			o	1749	17.32	39.29	29.50	9.43	3.60	0.40	0.34	0.06	0.06
	Anandamath	prob.	c	463	22.46	45.14	22.68	8.42	1.08	0.22			
			o	646	11.76	37.47	35.29	11.30	2.63	0.93	0.62		
		syst.	c	246	21.54	50.41	21.14	3.25	3.25	-	-	0.41	
			o	555	14.77	38.74	30.46	11.35	3.06	0.90	0.36	0.36	
		pooled	c	709	22.14	46.97	22.14	6.63	1.83	0.14	-	0.14	
			o	1201	13.16	38.05	33.05	11.32	2.83	0.92	0.50	0.17	
	Devi Choudhurani	prob.	c	469	22.60	53.09	19.83	3.84	0.64				
			o	705	16.60	38.16	34.88	7.80	1.99	0.43	-	-	0.14
		syst.	c	358	26.54	49.99	18.16	4.19	1.12				
			o	475	16.84	40.00	34.95	5.89	2.11	-	0.21		
		pooled	c	827	24.30	51.75	19.11	3.99	0.85				
			o	1180	16.69	38.91	34.93	7.03	2.03	0.25	0.08	-	0.08

Table 4.3 : (Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bankim-chandra	Rajsinha	prob.	c	530	17.74	46.22	24.90	8.87	1.89	0.19	0.19		
			o	893	15.68	36.62	30.00	11.65	4.48	1.23	0.34		
		syst.	c	184	17.39	46.20	28.26	7.61	0.54				
			o	323	12.38	34.67	30.97	12.69	5.57	2.48	0.62	0.31	0.31
		pooled	c	714	17.65	46.22	25.77	8.54	1.54	0.14	0.14		
			o	1216	14.80	36.11	30.27	11.92	4.77	1.56	0.41	0.08	0.08
Rabindra-nath	Bouthakuranir Hat	prob.	c	674	22.11	44.06	25.37	6.38	1.78	0.30			
			o	918	13.07	39.98	35.29	7.19	4.03	0.44			
		syst.	c	328	24.39	45.12	23.48	5.79	1.22				
			o	499	11.82	35.08	42.08	6.41	3.61	1.00			
		pooled	c	1002	22.85	44.41	24.75	6.19	1.60	0.20			
			o	1417	12.63	38.24	37.69	6.92	3.88	0.64			
	Rajarsi	prob.	c	413	23.73	46.50	21.79	6.05	1.45	0.48			
			o	1219	11.98	42.25	33.31	9.02	2.54	0.57	0.25	-	0.08(10)
		syst.	c	164	18.29	46.95	29.88	3.66	0.61	0.61			
			o	525	11.81	40.18	34.29	10.67	2.10	0.76	0.19		
		pooled	c	577	22.18	46.63	24.09	5.37	1.21	0.52			
			o	1744	11.93	41.62	33.60	9.52	2.41	0.63	0.23	-	0.06

Table 4.3 (Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Chokher Bali	syst.	c	380	22.11	48.16	22.37	6.84	0.26	0.26			
			o	938	12.47	43.28	33.69	8.32	1.81	0.32	0.11		
	Gora	prob.	c	353	20.11	54.68	20.11	4.25	0.85				
			o	536	12.50	42.92	32.84	8.21	3.35	0.18			
		syst.	c	682	20.82	55.57	16.72	5.57	1.32				
			o	1142	12.87	42.55	32.14	8.84	2.98	0.44	0.09	0.09	
		pooled	c	1035	20.58	55.27	17.87	5.12	1.16				
			o	1678	12.75	42.67	32.36	8.64	3.10	0.36	0.06	0.06	
	Chaturanga	prob.	c	329	20.36	48.03	27.06	3.95	0.30	0.30			
			o	1129	15.15	45.08	30.38	6.73	2.30	0.09	0.27		
		syst.	c	193	18.13	44.56	30.57	5.18	1.04	0.52			
			o	661	17.55	45.83	27.69	6.96	1.82	0.15			
		pooled	c	522	19.54	46.75	28.35	4.41	0.57	0.38			
			o	1790	16.03	45.36	29.39	6.82	2.12	0.11	0.17		
	Ghare Baire	prob.	c	530	20.38	61.13	13.77	3.40	0.75	0.38	0.19		
			o	1371	20.79	55.72	16.78	5.32	1.24	0.15			

Table 4.3 : (Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindra-nath	Sheser Kavita	prob.	c	338	23.67	52.06	17.46	5.62	0.89	0.30			
			o	397	15.11	56.18	20.15	5.29	3.02				0.25(11)
		syst.	c	700	18.43	55.29	19.86	5.71	0.57	0.14			
			o	584	17.98	52.40	19.86	6.51	1.71	1.20	0.17	0.17	
		pooled	c	1038	20.13	54.24	19.08	5.68	0.67	0.19			
			o	981	16.82	53.93	19.99	6.01	2.24	0.71	0.10	0.10	0.10(11)
	Yogayog	syst.	c	366	23.50	57.38	16.12	2.19	0.55	-	0.27		
			o	821	17.90	54.69	15.96	8.65	2.19	0.61			
Pramatha Chaudhury	Char-Yari Katha	prob.	c	298	25.50	58.72	12.08	3.36	0.34				
			o	574	21.25	55.93	15.33	5.57	1.05	0.70	0.17		
Saratchandra	Pallisamaj	prob.	c	373	31.10	52.82	12.60	2.14	1.34				
			o	517	12.38	44.88	34.04	5.22	2.71	0.39	0.19	0.19	
	Pather Dabi	prob.	c	353	25.50	52.11	18.70	3.12	0.57				
			o	462	16.45	39.17	35.28	7.36	1.30	0.22	0.22		
Bibhuti-bhusan	Pather Panchali	prob.	c	299	27.77	56.19	12.37	3.34	0.33				
			o	623	14.13	45.75	28.89	7.70	3.53				
		syst.	c	370	28.38	55.95	12.97	2.43	0.27				
			o	1260	13.57	46.98	29.29	7.78	1.90	0.40	0.08		
		pooled	c	669	28.10	56.05	12.71	2.84	0.30				
			o	1883	13.75	46.58	29.16	7.75	2.44	0.27	0.05		

Table 4.3 : (Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Bibhuti-bhusan	Aparajita	syst.	c	499	24.85	55.11	16.03	3.21	0.80				
			o	1395	15.99	44.81	29.46	6.52	2.65	0.43	0.07	-	0.07(10)
	Devayan	prob.	c	559	23.08	56.53	13.77	5.37	1.07	0.18			
			o	372	16.13	55.38	19.35	7.26	1.88				
		syst.	c	1194	19.85	56.20	18.09	4.94	0.75	0.17			
			o	1051	18.74	55.57	16.94	6.95	1.52	0.19	0.09		
		pooled	c	1753	20.88	56.30	16.71	5.08	0.86	0.17			
			o	1423	18.06	55.51	17.57	7.03	1.62	0.14	0.07		
Jajabar	Dristipat	prob.	c	100	25.00	52.00	18.00	4.00	1.00				
			o	672	16.07	43.60	25.30	11.31	3.27	-	0.30	0.15	
		syst.	c	181	23.76	51.94	16.57	6.63	1.10				
			o	1410	15.46	45.47	24.89	9.72	3.19	1.13	0.07	0.07	
		pooled	c	281	24.20	51.96	17.08	5.69	1.07				
			o	2082	15.66	44.86	25.02	10.23	3.22	0.77	0.14	0.10	
	Janantik	prob.	c	176	19.32	59.66	14.77	5.68	-	0.57			
			o	514	14.79	50.38	22.96	8.37	2.72	0.78			

Table 4.3 : (Contd.)

author	work	type of sample	class of words (c/o)	no. of sample words	percentage of words by length in syllables								
					1	2	3	4	5	6	7	8	9-
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Muztaba Ali	Chacha-Kahini	prob.	c	139	19.42	58.99	17.27	2.88	1.44				
			o	639	16.12	55.71	20.97	5.63	1.25	0.16	0.16		
	Deshe Videshe	prob.	c	281	21.00	55.51	18.51	3.56	1.42				
			o	510	17.06	55.68	17.84	7.65	1.57	-	0.20		
Rabindranath	Kabuliwalla	syst.	c	82	18.29	53.66	20.73	4.88	1.22	1.22			
			o	697	13.06	39.88	33.43	9.61	2.44	1.29	0.29		
	Kshudhita Pasan	syst.	c	51	35.29	37.26	25.49	0	1.96				
			o	1141	10.87	44.08	31.29	9.47	2.80	0.96	0.09	0.35	0.09(10)
	Laboratory	syst.	c	845	23.91	56.09	14.44	3.55	1.66	0.24	0.12		
			o	383	13.32	53.52	22.72	8.88	0.78	0.52	0.26		

4.2.5. As regards the sampling properties of these estimates, it seems that the situation is not at all happy for the estimates p_c shown in Table 4.1. These estimates have the same ratio form as p_r of Chapter 2, para 2.4.1, [note that r is an integer, but c is not]. For if n_i is the number of words on the ith sample line or cluster of words, and $n_i^{(c)}$ the number out of these falling under conversation

$$p_c = \frac{\sum_i n_i^{(c)}}{\sum_i n_i}$$

where $\sum_i$ denotes summation over all the k clusters (lines) in the sample or subsample. Now k is much greater than 30 for the combined samples, and at least 25 for the subsamples (in most cases). Also C. V. $(\bar{n}_i)$ has been seen in Chapter 2, Section 2.5, to be below 0.1 even for the subsamples. But the C.V. of the sample average of $n_i^{(c)}$ over clusters would be very large and not below 0.1 in most cases.

4.2.6. No actual calculation of C.V. is needed for such examination. A simple model indicates the situation. Suppose all clusters (lines) have the same number (m) of words, and the sample cluster is either wholly conversational with probability π or wholly non-conversational with probability $1-\pi$. This is not very unrealistic for conversational and other words tend to occur in long runs, and consequently, sample lines are usually either wholly conversational or wholly otherwise. The C.V. of the sample average of $n_i^{(c)}$ would then be given by

$\sqrt{\frac{1-\pi}{k\pi}}$, where k is the number of sample clusters. This will be less than 0.1, as required, if $k > 100$ when $\pi = 0.5$, if $k > 400$ for $\pi = 0.2$, and if $k > 900$ when $\pi = 0.1$. Clearly, the large sample properties are not possessed by most of the estimates p_c given in Table 4.1.

4.2.7. As regards the estimates $\bar{x}_c$ and $\bar{x}_o$, they should have properties similar to those of the overall averages $\bar{x}$, excepting that they are based on smaller sample sizes. In view of the findings of Chapter 2, Sections 2.4 - 2.5, it may be presumed that the estimates of $\bar{x}_c$ and $\bar{x}_o$ possess the large sample properties of ratio estimates at least to a rough approximation. The approximation is poor for some estimates $\bar{x}_c$ where p_c is very small, so that the underlying sample size is not quite adequate.

4.2.8. It seems to be unnecessary to test the validity of the systematic sample estimates.

4.3.1. Percentage of Conversational matter (p_c) : We may now consider the estimates p_c presented in Table 4.1. The subsample estimates diverge greatly, so the combined estimates[1] have wide margins of error. They may deviate by more than 5% from the true values for many of the works and by more than 10% for some works with smaller sample sizes (e.g., "Shakuntala"). This is not unexpected, and the problem was met in a different form in para 4.2.6. above. Assuming the model described there, the standard error of the estimate p_c would be given by

---

1/ 'Pooled' estimates, where both types of samples are available.

$\sqrt{\frac{\pi(1-\pi)}{k}}$. Actually, p_c is like an estimate of a binomial proportion estimated from a random sample of k units only. This explains the wide margins of sampling error. But the situation can hardly be helped. If $\pi = 0.5$, and if one wants to make the width of the confidence interval for π (with confidence coefficient 95%) only 2% (say) one should sample 2500 lines, approximately, which is evidently too much.

4.3.2. Anyway, it is fortunate that even these rough estimates p_c do lead to interesting conclusions. Consider first the following statement showing the distribution of the different works over levels of p_c. The estimates being subject to wide margins of error, some works may in reality belong to levels higher or lower than those shown here.

estimated percentage of conversational matter (p_c)	names of works
0 - 5	Kshudhita Pasan (4.28%)
5 - 10	....
10 - 15	Kabuliwalla, Dristipat
15 - 20	Visavriksha, Chacha Kahini
20 - 25	Rajarsi, Chaturanga
25 - 30	Durgeshnandini, Chokher Bali, Ghare Baire, Pather Panchali, Aparajita, Janantik
30 - 35	Kapalkundala, Krishnakanter Will, Yogayog, Char-Yari Katha
35 - 40	Anandamath, Rajsinha, Gora, Deshe Videshe
40 - 45	Devi Choudhurani, Bouthakuranir Hat, Pallisamaj, Pather Dabi
45 - 50	Sitar Vanavas
50 - 55	Shakuntala, Sheser Kavita
55 - 60	Devayan
60 - 65	....
65 - 70	Laboratory (68.81%)

4.3.3. The table shows continuous variation in p_c from near zero to about 70%; and the percentages are exactly or practically zero for essays or collections of essays like "Birbaler Halkhata" not included here. This remarkable variation cannot be easily explained. Both the extreme positions are occupied by short stories; but even among novels in standard style one finds p_c as low as 16.81% ("Visavriksha") and as high as 41.91% ("Pallisamaj").

4.3.4. For a good number of works, as already noted, p_c does not tell the whole story. "Kabuliwalla", "Chaturanga", "Dristipat", "Deshe Videshe" and "Chacha Kahini" are written as reminiscences of the author, who is mentioned in the first person. In "Ghare Baire", "Kshudhita Pasan" and "Char-Yari Katha", on the other hand, the different divisions are like reminiscences of different characters. In all these works, conversations are distinguishable from other matter, but the latter are also somewhat akin to conversation.

4.3.5. Appreciable within author differences are noticed in several cases, and most of these seem to be significant.[1/] Compare, for instance, "Pather Panchali" or "Aparajita", on the one hand, with "Devayan", on the other [vide Section 4.1 supra]. "Visavriksha" and "Krishnakanter Will" are considered to be closely similar in subject-matter, but they show very different values of p_c and the difference

1/ For tests of significance, we used the two-sample t-test, there being four subsample estimates for each work. But this test may not be strictly valid here. The sign test suffices for practical purposes. For each subsample, we consider the signs of differences between the two estimates for two works. If all four signs are alike, we may take the difference as significant, although the level of significance is $6\frac{1}{4}\%$, even if the test is one-sided. Where both types of samples are available, we may examine the eight subsample estimates to be on surer ground.

is evidently significant. The position is similar for "Rajarsi" and "Bouthakuranir Hat", but the significance is not so very evident here. "Chacha Kahini" and "Deshe Videshe" can be added to this list, and also "Dristipat" and "Janantik"; but the two works of either pair are not strictly comparable, particularly those of the second pair.[1/] In most of these cases, the within author difference between the overall $\bar{x}$'s is partly explained by the corresponding differences in the precentages p_c; for the $\bar{x}_c$'s and the $\bar{x}_o$'s are closer for the two works than the overall averages $\bar{x}$.

4.4.1. Observations of $\bar{x}_c$ and $\bar{x}_o$: The relevant figures are given in Table 4.2. Although the sampling errors are larger than those for the overall averages $\bar{x}$, some important conclusions can still be reached from these estimates for $\bar{x}_c$ and $\bar{x}_o$.

4.4.2. It is obvious that $\bar{x}_o$ is significantly larger than $\bar{x}_c$. The combined estimates show $\bar{x}_o > \bar{x}_c$ in all the rows of the table. There is no doubt, therefore, that, on the whole, the difference is real. But as regards individual works, the difference cannot be declared significant in all the cases. The inequality $\bar{x}_o > \bar{x}_c$ does not hold between the subsample estimates in all the cases. True, $\bar{x}_c - \bar{x}_o$ is seldom appreciably above zero. There are only two notable examples, one in the subsample 2 of the probability sample from "Ghare Baire", and the other in subsample 1 of the systematic sample from "Chaturanga", and even these

1/ More serious differences can be pointed out if one compares the different works by Tagore or Bamkimchandra, without regard to differences in subjectmatter or age of author.

may easily be ascribed to sampling errors, for the number of conversational words in these two cases are only 102 and 45 respectively, and the true difference $\bar{x}_o - \bar{x}_c$ seems to be small for both the works. But in any case, if one of the four subsamples shows the inequality reversed, it cannot be said that $\bar{x}_o > \bar{x}_c$ significantly for that particular work. If both types of samples are available, one might consider the eight signs of $\bar{x}_o - \bar{x}_c$ together, although the two samples might have very different sizes; and one may declare significance if at least seven of the eight signs are positive.[1/]

4.4.3. Most of the works show $\bar{x}_o$ significantly larger than $\bar{x}_c$. The only clear exception is "Ghare Baire". Also, the significance is not clear in several cases, viz., "Chaturanga", "Sheser Kavita", "Devayan", "Chacha Kahini" and "Deshe Videshe", apart from two short stories ("Kabuliwalla" and "Kshudhita Pasan") where the sample size for conversational words is very small. It is interesting to note that works named here include most of those in which the non-conversational matter is akin to conversation (vide para 4.3.4).

4.4.4. It may be interesting to digress for comparing the entire length distributions of the two classes of words. Table 4.4 presents the distributions for two selected works, merging the four subsamples into two half-samples. Only the probability sample for "Krishnakanter Will" has been utilised. Table 4.5 presents the same distributions in

1/ Apart from the sign test, we also applied Student's t-test on the subsample differences $\bar{x}_o - \bar{x}_c$, for examining whether the overall difference is significantly above zero.

terms of decile group averages. Fig. 4.1 represent the estimates of Table 4.5 in the form of fractile graphs. The separation between the two classes of words is clearly significant for "Krishnakanter Will", at least from the sixth decile group upwards. For "Ghare Baire", on the other hand, the separation is noticeable only at the three highest decile groups, where the divergence seem to be significant.

4.4.5. Fig. 4.2 shows the values of $\bar{x}_o$ and $\bar{x}_c$ for different works, each work being represented by a point in the scatter diagram. The name of the work is shown against each point. The combined (all subsample) estimates were used; where two types of samples were available, the pooled estimate was used. All the points in the diagram are not equally reliable. For instance, the point for "Kapalkundala" is based on only 149 conversational words and 344 other words, while that for "Anandamath" is based on 685 conversational words and 1201 other words. Even so, the graph is illuminating and leads to the following observations. (One should keep the subsample estimates of Table 4.2 in mind while drawing conclusions from Fig. 4.2.)

4.4.6. On the whole, the two averages are fairly correlated in the positive sense. This means, in plain words, a work with a higher value of $\bar{x}_o$ is likely to have a higher value of $\bar{x}_c$ as well. But there are large deviations from the general pattern (i.e., the regression line implicit in Fig. 4.2).

Table 4.4: Percentage distribution of words by length in syllables, estimated from probability samples of words from two selected works in Bengali prose, separately for conversational and other words

length (no. of syllables)	Krishnakanter Will						Ghare Baire					
	percentage of conversational words			percentage of other words			percentage of conversational words			percentage of other words		
	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	20.23	26.64	23.48	18.48	16.49	17.46	20.78	20.00	20.38	21.95	19.47	20.79
2	53.57	46.33	49.90	40.84	38.99	39.89	61.18	61.10	61.14	55.55	55.92	55.72
3	21.03	22.01	21.53	28.69	29.74	29.22	14.51	13.09	13.77	16.32	17.29	16.78
4	3.57	4.25	3.91	9.24	9.40	9.32	2.75	4.00	3.40	4.94	5.76	5.32
5	1.60	0.77	1.18	2.11	4.62	3.40	-	1.45	0.75	1.10	1.40	1.24
6				0.32	0.46	0.39	0.39	0.36	0.37	0.14	0.16	0.15
7				0.32	0.15	0.24	0.39		0.19			
8					-	-						
9					0.15	0.08						
total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
average length	2.13	2.06	2.09	2.38	2.50	2.44	2.03	2.07	2.05	2.08	2.14	2.11
no. of words	252	259	511	617	649	1266	255	275	530	729	642	1371

Table 4.5 · Decile group averages of word-length in syllables, estimated from probability samples of words from two selected works in Bengali prose, separately for conversational and other words

fractile group (per cent)	Krishnakanter Will						Ghare Baire					
	conversational words			other words			conversational words			other words		
	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.	h.s.1	h.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
0 - 10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10 - 20	1.00	1.00	1.00	1.15	1.35	1.25	1.00	1.00	1.00	1.00	1.05	1.00
20 - 30	1.98	1.34	1.65	2.00	2.00	2.00	1.92	2.00	1.96	1.80	2.00	1.92
30 - 40	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
40 - 50	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
50 - 60	2.00	2.00	2.00	2.07	2.45	2.26	2.00	2.00	2.00	2.00	2.00	2.00
60 - 70	2.00	2.00	2.00	3.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00
70 - 80	2.62	2.70	2.66	3.00	3.00	3.00	2.00	2.00	2.00	2.25	2.46	2.35
80 - 90	3.00	3.00	3.00	3.20	3.48	3.34	2.80	2.89	2.85	3.00	3.00	3.00
90 -100	3.68	3.58	3.63	4.37	4.66	4.53	3.55	3.80	3.68	3.76	3.90	3.82
0 -100	2.13	2.06	2.09	2.38	2.50	2.44	2.03	2.07	2.05	2.08	2.14	2.11
no. of words	252	259	511	617	649	1266	255	275	530	729	642	1371

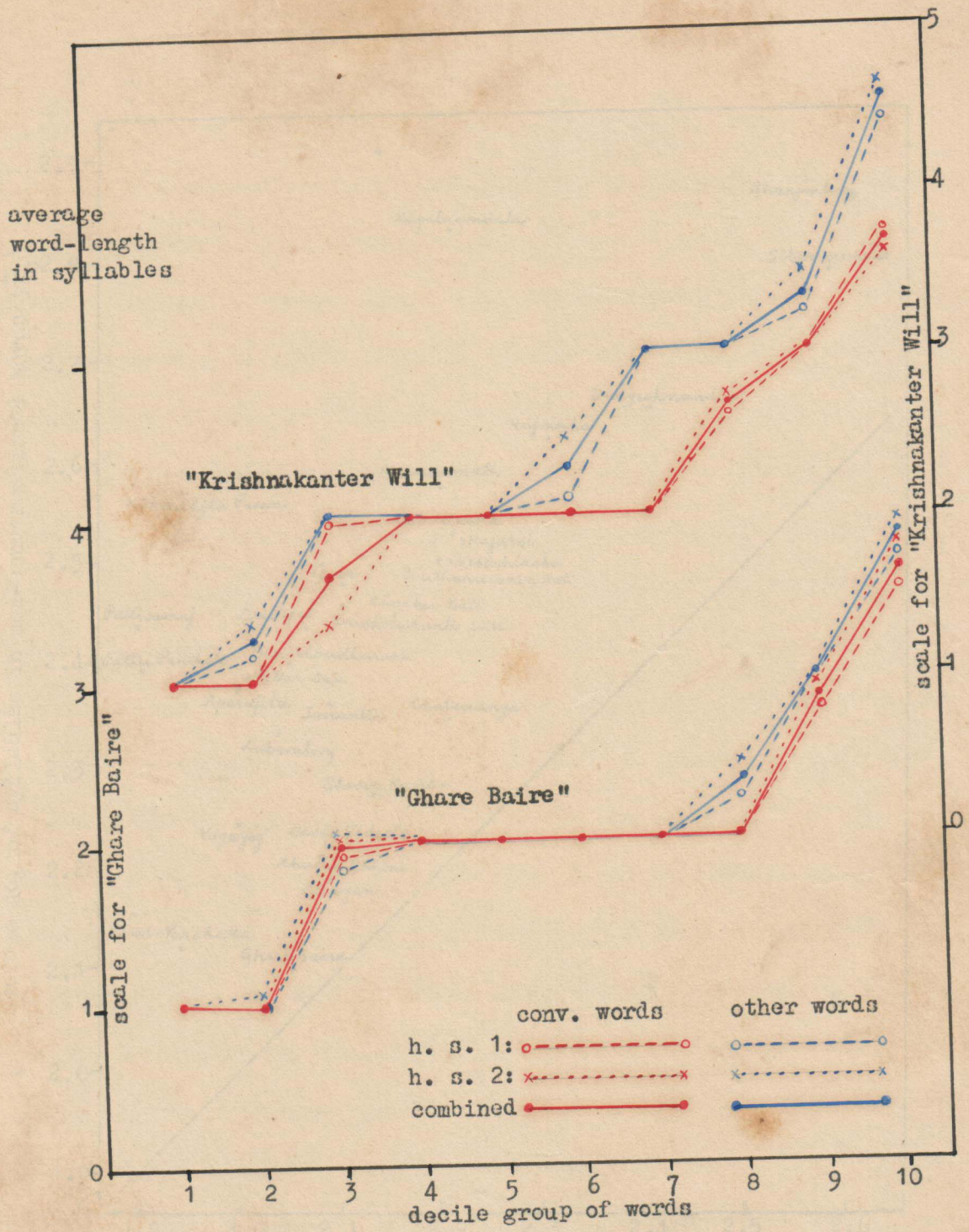


Fig. 4.1: Fractile graph for word-length in syllables, estimated from probability samples of words from "Krishnakanter Will" and "Ghare Baire", separately for conversational words and other words [Vide Table 4.5].

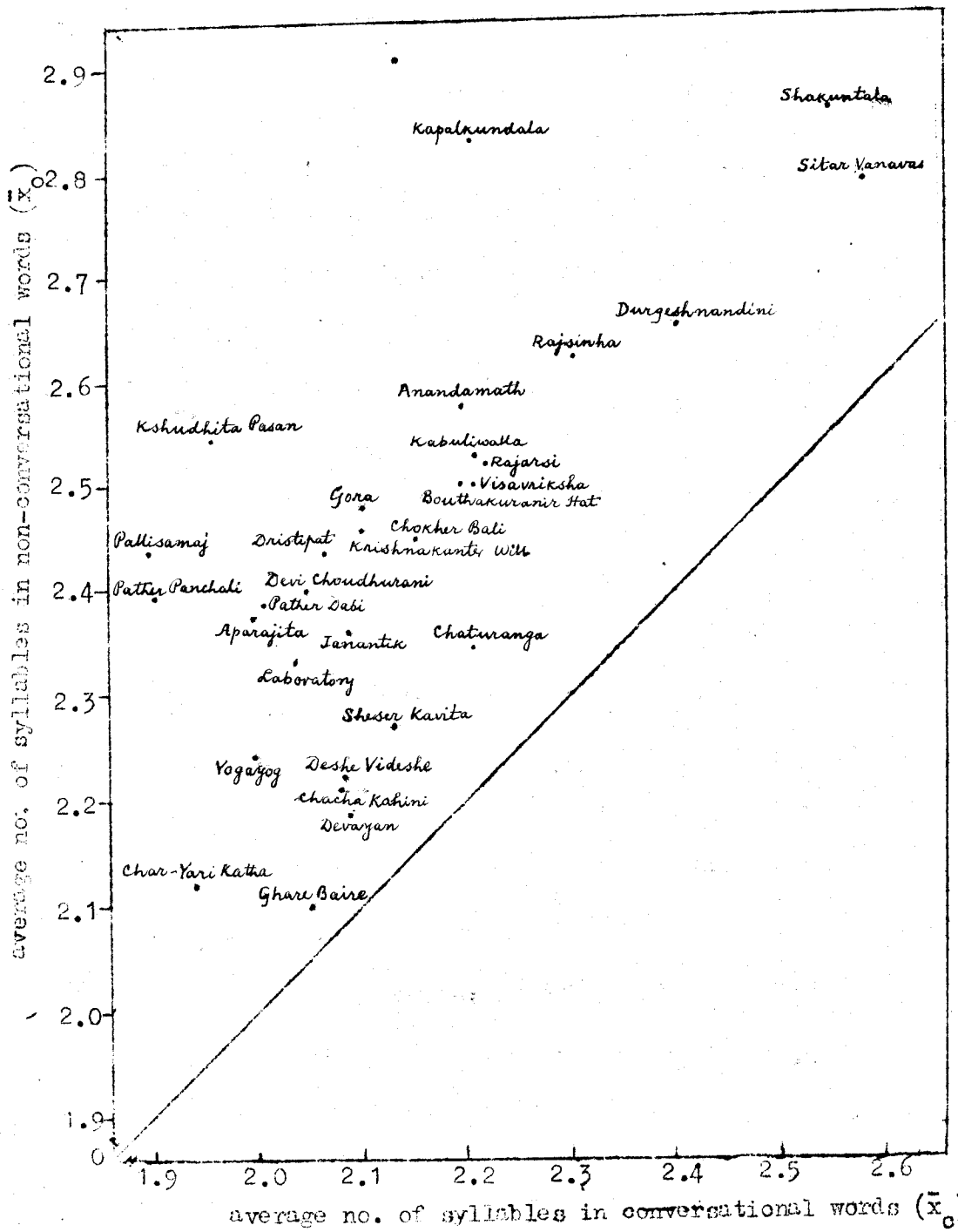


Fig.4.2: Scatter diagram showing average lengths of conversational words and other words, estimated for selected works in Bengali prose, on the basis of probability and/or systematic samples of words (vide Table 4.2).

4.4.7. The difference $\bar{x}_o - \bar{x}_c$ seems to be small, about 0.1 or 0.15, for a number of works written entirely in the colloquial style ("chalita bhasa"), viz., "Ghare Baire", "Sheser Kavita", "Char-Yari Katha", "Devayan", "Chacha Kahini" and "Deshe Videshe". Among these, the difference is a little large (0.18) for "Char-Yari Katha" and rather small (0.06) for "Ghare Baire". "Chaturanga", written in "Sadhu Bhasa" (Chaste Style) throughout is seen to belong to this group; this work, as stated earlier, uses a de facto colloquial style, only the verbs and pronouns having the chaste form. "Yogayog", written in the colloquial style shows a somewhat higher value of $\bar{x}_o - \bar{x}_c$, but the difference may not be significant. It may be pointed out that most of the works mentioned in this para are written as speeches or thoughts of the author or some leading characters of the story.

4.4.8. The other three works written in the colloquial style are "Laboratory", "Dristipat" and "Janantik"; these seem to have larger differences $\bar{x}_o - \bar{x}_c$ than the works mentioned in the foregoing paragraph. "Dristipat" and "Janantik" are evidently written in a different style; but the anomalous position for "Laboratory" cannot be explained in this way, and must be largely due to sampling errors.

4.4.9. Five works, viz., "Gora", "Pallisamaj", "Pather Dabi", "Pather Panchali" and "Aparajita", employ the colloquial style in conversations but the chaste style elsewhere. The difference $\bar{x}_o - \bar{x}_c$ seems to be larger for these works, between 0.35 and 0.5. The differences $\bar{x}_o - \bar{x}_c$ for the remaining works wholly in chaste style (e.g., "Shakuntala", "Sitar Vanavas") seem to be nearly equal to or a little less than

those in the five works using the chaste style outside conversations but the colloquial style within conversations.

4.4.10. The point for "Kapalkundala" is not very reliable, but it seems that it has the highest difference $\bar{x}_o - \bar{x}_c$ (viz., 0.63). The $\bar{x}_c$-values for "Kshudhita Pasan" and "Kabuliwalla" are even more unreliable.

4.4.11. Fig. 4.2 reveals that the works of Vidyasagar, Bamkimchandra and Tagore (upto "Chokher Bali")— all wholly in the chaste style — fall around a straight line, on which $\bar{x}_o$ is greater than $\bar{x}_c$ by roughly 0.35. This is a part of the declining trend in $\bar{x}$ (vide Chapter 3, Section 3.2) defined in a more detailed manner. Thereafter occurred a major change : the colloquial style began to be used in conversational matter, though not elsewhere. The five works employing such "mixed" styles ("Gora", "Pallisamaj", "Pather Dabi", "Pather Panchali" and "Aparajita") seem to continue the declining linear trend described earlier. Or rather, since the non-conversational matter retained the chaste style, $\bar{x}_o$-values show smaller decrease than $\bar{x}_c$-values during this period.

4.4.12. Fig. 4.2 also shows the effect of the movement, championed by Pramatha Choudhury and Rabindranath Tagore, for using the colloquial language throughout the text, both in conversations and elsewhere. While the $\bar{x}_c$-values retained their order of magnitude — the colloquial style was already there in the conversations — the $\bar{x}_o$-values fell greatly. So, most of the works entirely in colloquial style fall

around another line with $\bar{x}_o - \bar{x}_c$ only about 0.1 or 0.15. The exceptional cases have already been mentioned in paras 4.4.7-8.

4.4.13. In para 4.3.5 above, we mentioned some pairs of more or less similar works showing noticeable within author differences in p_c. Such differences furnish partial explanations of the corresponding differences in the overall $\bar{x}$'s. The finding mentioned in the introductory section of this chapter may be recalled here. If p_c were the same as in "Devayan", the value of $\bar{x}$ would have been 2.129 for "Pather Panchali" and 2.166 for "Aparajita", as against 2.138 for "Devayan".

4.4.14. We now examine the position for the other pairs also, viz., for "Visavriksha" and "Krishnakanter Will", for "Rajarsi" and "Bouthakuranir Hat" and for "Dristipat" and "Janantik". The relevant figures for these works are reproduced below.

work	p_c	$\bar{x}_c$	$\bar{x}_o$	$\bar{x}$	adjusted* $\bar{x}$
Visavriksha	16.81	2.227	2.505	2.459	2.419
Krishnakanter Will	30.76	2.103	2.459	2.350	2.399
Bouthakuranir Hat	41.42	2.199	2.530	2.393	2.448
Rajarsi	24.86	2.184	2.521	2.437	2.381
Dristipat	11.89	2.075	2.439	2.395	2.346
Janantik	25.51	2.091	2.362	2.293	2.330

* See para 4.4.15 for explanation

4.4.15. If "Visavriksha" had the same p_c as "Krishnakanter Will", the present values of $\bar{x}_c$ and $\bar{x}_o$ would give $\bar{x}$ = 2.419, which is closer to the $\bar{x}$ for "Krishnakanter Will", than the actual $\bar{x}$ (2.459) for "Visavriksha". Similarly, if "Krishnakanter Will" had p_c = 16.81%, as

in "Visavriksha", its $\bar{x}$ would have been 2.399. Thus, variation in p_c has increased the difference in $\bar{x}$ between these works. As regards "Rajarsi" and "Bouthakuranir Hat", the observed difference in $\bar{x}$ is almost entirely due to unequal p_c's, for the $\bar{x}_c$'s and the $\bar{x}_o$'s are nearly equal. Thus, if "Rajarsi" had the same p_c as "Bouthakuranir Hat", its $\bar{x}$ would come to 2.381 and similarly, if "Bouthakuranir Hat" had the same p_c as "Rajarsi" it would have $\bar{x}$ = 2.448. The difference in $\bar{x}$ between "Dristipat" and "Janantik" is also partly due to differences in p_c. For if, "Dristipat" had the same weightage of conversation as "Janantik" it would show $\bar{x}$ = 2.346, and if "Janantik" possessed the same percentage p_c as "Dristipat", its $\bar{x}$ would have risen to 2.330.*

4.4.16. The distinction between conversational and other words does not seem to have been made in earlier studies on language. It should be abundantly clear that this simple distinction gives a far clearer picture than can be obtained without it. Thus, while $\bar{x}_c$ seems to be smaller for "Pather Panchali" than for "Devayan", $\bar{x}_o$ seems to be larger for the former than for the latter. Some work done on the English novel, "Pride and Prejudice", is reported in Appendix 3. The difference $\bar{x}_o - \bar{x}_c$ is appreciable and significant in "Pride and Prejudice" also. The contrast between conversational matter and other matter in Bengali prose will become further apparent in Chapter 5, Section 5.5 , where some short passages will be examined for assessing how far the series of word-lengths is random in the statistical sense.

---

* These hypothetical values are shown in the last column of the tabular statement in para 4.4.14.

Chapter 5 : Randomness of the Wordlength Series

5.1.1. Introductory: An analogue of a time series is obtained if all words of a given text are replaced by the corresponding word-lengths and these lengths are read in the normal reading order. Word-length may be measured in either syllables or phonemes or letters. The randomness of this series was discussed by Fucks (1954) who measured word-length in syllables. Fucks distinguished between the correlation between lengths of two or more consecutive words, termed Nahordnung, and the correlation between lengths of two or more words which are not consecutive (Fernordnung). It is simpler to think in terms of autocorrelations of the word-length series, which may be used for studying both Nahordnung and Fernordnung.

5.1.2. Fucks suggested many methods for studying such correlations[1]. Of greater interest are his empirical findings, based on six works in German and one in English (viz., Shakespeare's "Othello"). Fucks did not mention any sample sizes - the same is true of Fucks (1952, 1955) also. Presumably he based his figures on complete counts, in which case the sampling errors may be taken to be negligible, if not zero[2]. Fucks showed that the autocorrelation

---

1/ Some of these are of dubious value, e.g., the measurement of skewness of the joint distribution of consecutive word-lengths, or the estimation of characteristic distance (Fucks, 1954, p.131).

2/ Vide Section 2.8 of Chapter 2 in this connection. Results based on complete counts cannot be strictly regarded as absolute, that is, without any sampling errors associated with them. The size of the complete count may therefore be given to give a rough idea about reliabilities.

coefficient of first order (r_1) is nearly zero for all the German and English works: Actually the values range from - 0.065 to + 0.013, four being negative, and three positive, of which one is 0.002. The joint distributions of lengths of two consecutive words were also presented for all the works. These tables showed that the lengths of consecutive words are approximately independent in the statistical sense: The entropies of these bivariate distributions were nearly double of the entropies of the corresponding marginal distributions.

5.1.3. The randomness of the series of word-lengths seems to be of considerable statistical interest, and will be examined in this chapter from a number of angles. We first examine, in Section 5.2, the significance of some results already reported in Chapters 3 and 4; we point out, for example, that the existence of "patches" of shortish (conversational) and longish (non-conversational) words may give rise to positive autocorrelations in the word-length series, even if no autocorrelation exists within such patches. Sections 5.3 and 5.4 are concerned with the estimates of autocorrelation coefficients. The estimate of r_1 was obtained for sixteen works in Bengali prose, mostly fiction. All these works were covered in Chapter 3, and the same probability samples of words have been used for the purpose. For four out of these sixteen works, some autocorrelation coefficients of higher order were estimated, so as to reveal the shape of the correlogram. Word-length has been measured in syllables. Section 5.5 examines the autocorrelations etc., in some short

continuous passages, selected from two of these four works in a purposive manner; the presence of "patches" is revealed by this examination. Finally, Section 5.6 reports on some work on English, where word-length was measured in terms of letters. The value of r_1 is found to be -0.1 or -0.15 in the two works examined, which partly contradicts the finding of Fucks (1954) who examined only one work viz., Shakespeare's "Othello", and found the value of r_1 to be nearly zero (-0.02992).

5.1.4. It should be added that the randomness or otherwise of given texts may be examined in various other ways, for instance, by comparing the word-length distributions of different parts of the work. One interesting approach would be to compare the distributions found from speeches by different characters with that found from the non-conversational matter.

5.2.1 Some evidence of deviations from perfect randomness : Before presenting the statistical material having direct bearing on randomness of word-length series, it seems worthwhile to examine certain pieces of evidence available from Chapters 3 and 4 which throw some light on the problem.

5.2.2. Consider, first, the estimated standard errors $s_{\bar{x}}$ of the sample averages of word-length $\bar{x}$ given in column (11) of Table 3.1 of Chapter 3. One may compare them with the values of $\frac{s_x}{\sqrt{n}}$ where s_x is the s.d. of word-length [col.(12) of Table 3.1] and n the number of sample words [col. (5) of same table]. This $\frac{s_x}{\sqrt{n}}$ would

have been the standard error of $\bar{x}$, if the sample were unrestricted random. If the actual standard errors $s_{\bar{x}}$ are different from $\frac{s_x}{\sqrt{n}}$, the difference must be due to the intra-cluster, that is, intra-line correlations between lengths of words[1].

5.2.3. It is well-known that, if the clusters are of equal size, say m, and if n clusters are selected strictly at random, then the true standard error of sample mean $\bar{x}$ bears to $\frac{\sigma_x}{\sqrt{n}}$ the ratio $\sqrt{1 + (m-1)\bar{\rho}}$ . . Here σ_x is the population s.d. of x and $\bar{\rho}$ is the average of intercorrelations between different pairs of units on the same cluster. If the clusters have unequal sizes, as in the present case, and the selection of clusters is done at random, there is no such simple result showing the effect of intra-line correlations[2], but obviously the difference will depend on them and will be due to them.

5.2.4. One may now see that, on the average, the estimated standard error $s_{\bar{x}}$ is greater by about 10% than the value of $\frac{s_x}{\sqrt{n}}$. But the percentage difference varies appreciably among the different rows of Table 3.1: There is a whole frequency distribution from -10% to + 34%. Only 4 values are negative; so the overall average, roughly 10%, may be regarded as significant. Assuming that m =8, this gives an overall estimate of $\bar{\rho}$ of about 0.03.

1/ Both $s_{\bar{x}}$ and $\frac{s}{\sqrt{n}}$ are estimates with certain margins of error, which have not been calculated.

2/ There is one elegant result for the case where clusters are selected with probability proportional to size [Sukhatme, 1954, pp. 270-272].

Estimates of $\bar{\rho}$ are not very interesting, however; for $\bar{\rho}$ is only a certain average of r_1, r_2, r_3....., the autocorrelation coefficients of different orders, which should be studied separately. But it is important to note that, due to the presence of autocorrelations in the word-length series, the sampling errors of averages based on the probability (and systematic) samples described in Chapter 2 tend to be 10% larger than those of strictly random samples of words of the same size.

5.2.5. The difference between conversational and non-conversational words may also be recalled in this connection. It was pointed out in Chapter 4 (see also Section 5.5 below) that texts of fictions tend to show alternate patches of longish (non-conversational) and shortish (conversational) words, with the average word-lengths in the two types of patches differing by 0.1 to 0.5 syllables, which is a considerable difference in relative terms. This is, of course, a clear non-random feature, and significant. It is important to note here that this feature may give rise to, or exaggerate, positive autocorrelations r_1, r_2 etc., while the neighbouring word-lengths <u>within</u> conversational or <u>within</u> non-conversational passages may be <u>relatively</u> independent in the statistical sense : When all pairs of neighbouring words are considered together, the "between groups" covariance may produce a large effect /Goulden, 1939, pp.247-50/. This point could not be examined throughly in this study, but this may be a partial explanation of the positive values of the autocorrelations found in Section 5.4.

5.2.6. That systematic samples of words behave like probability samples to a high degree of approximation [Chapter 2, Section 2.7] also points to the nearly perfect randomness of the word-length series. Or rather, the series seems to be like a stationary stochastic process with the autocorrelations falling away rapidly and vanishing beyond a few lines at most; there is no trend or periodicity with periods comparable with the intervals used for systematic sampling in the present study. (Roughly speaking, the intervals were usually of the order of one page or so.) Otherwise the systematic samples could not agree so well with the probability samples for all the works even in respect of sampling errors.

5.3.1. The Estimation of Autocorrelation Coefficients: We may now consider the estimation of autocorrelation coefficients of word-length series. The first point to be emphasised is the distinction between the autocorrelations in comparatively short continuous passages and the overall autocorrelations for entire works based on complete counts or on samples. The latter autocorrelations are evidently of greater interest and will receive greater emphasis in the present Chapter. But a few short passages from certain works will also be studied to see the within passage autocorrelation [vide Sections 5.5 and 5.6] This could show whether and how far the between patches variation, arising either out of the difference between conversational and other words or otherwise, makes the overall autocorrelation appreciably different from the within passage autocorrelation.

5.3.2. In the work concerned with Bengali prose, sixteen works were covered for estimation for r_1 (vide Tables 5.1 and 5.2), and four out of these for the estimation of r_2, r_3, r_4 and r_7 (vide Tables 5.3 to 5.7). Only the probability samples of words were utilised. From each of these works, it may be recalled, 100 sample lines were selected at random to get the probability samples of words; the four subsamples each comprised words falling on 25 sample lines (vide Chapter 2, Section 2.5).

5.3.3. Given the probability sample of words from any particular work, the first task was to prepare a two-way distribution by length of all pairs of words having the specified number of words (s) intervening between them (s =0,1,2,3 and 6). Obviously $n_i - s - 1$ pairs could be formed from the n_i words falling on the ith sample line. But in order that all possible word-pairs of the text, with s words between them, might be properly represented in the sample of pairs, the s + 1 words of the line following the sample line were also used, so that each word of the sample line became, in turn, the preceding word of one pair.

5.3.4. The two-way distributions of lengths of consecutive words are presented in Table 5.1; those for words at gap 1 in Table 5.3; those for words at gap 2 in Table 5.4; those for wards at gap 3 in Table 5.5; and those for words at gap 6 in Table 5.6. The sub-sample distributions are not presented for considerations of space. The autocorrelation coefficients are shown by sub-samples in Tables 5.2 and 5.7. Nevertheless, the two-way distributions point to the fact that

the lengths of neighbouring words are very nearly independent in the statistical sense.

Table 5.1: Joint distributions of lengths of consecutive words, based on probability samples of words from sixteen works in Bengali prose.

(a) Shakuntala

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	6	7	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	10	35	24	10	1	1		81
2	27	67	92	37	12	5	1	241
3	33	75	86	22	15	4		235
4	10	37	28	11	8	1		95
5	2	9	15	4	1			31
6		3	6	2		1		12
7		1						1
total	82	227	251	86	37	12	1	696

(b) Sitar Vanavas

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	6	7	8	10	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	21	31	41	11	4	3	1		1	113
2	34	81	74	25	17	3		1		235
3	40	82	89	26	14	3	3			257
4	11	29	30	9	5	2				86
5	4	10	19	4	3	1				41
6	1	4	4	1	1	1				12
7		3	1							4
8			1							1
10			1							1
total	111	240	260	76	44	13	4	1	1	750

Table 5.1: (Contd.)

(c) Durgeshnandini

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	6	7	8	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	17	36	24	9	1	1	1		89
2	41	79	57	22	5	2		1	207
3	19	66	70	23	12				190
4	12	17	19	5	2	1	1		57
5	3	7	5	6		3			24
6		4	2	1					7
7	1		1						2
8		1							1
total	93	210	178	66	20	7	2	1	577

(d) Visavriksha

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	6	7	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	20	41	31	5		1	1	99
2	41	93	81	24	6	3		248
3	28	72	47	23	6			176
4	8	19	24	7	4			62
5		7	9	1		2	1	20
6		1	3					4
7		1		1				2
total	97	234	195	61	16	6	2	611

Table 5.1 : (Contd.)

(e) Gora

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	20	75	30	11	2			138
2	72	199	118	26	7	1	1	424
3	29	110	86	14	7	1		247
4	9	32	12	3	3			59
5	3	9	4	2	2			20
6			1					1
total	133	425	251	56	21	2	1	889

(f) Sheser Kavita

length of preceding word in syllables	length of following word in syllables						total
	1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	30	73	30	7			140
2	80	216	74	18	10	1	399
3	26	73	27	10	3		139
4	8	26	4	2			40
5	1	8	4	1	1		15
6		1					1
11	1						1
total	146	397	139	38	14	1	735

Table 5.1 : (Contd.)

(g) Char-Yari Katha

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	47	109	31	7	4			198
2	110	288	63	30	3	2	1	497
3	26	72	20	3	1	1		123
4	7	23	7	4		1		42
5	2	4	1					7
6		3	1					4
7		1						1
total	192	500	123	44	8	4	1	872

(h) Birbaler Halkhata

length of preceding word in syllables	length of following word in syllables								total
	1	2	3	4	5	6	7	9	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	62	80	49	21	7	1			220
2	96	208	106	31	13	1	2	1	458
3	40	126	54	12	10	4	1		247
4	11	29	23	6	1		1		71
5	5	14	11	1	2				33
6	2	4				1			7
7	2	1	1						4
9			1						1
total	218	462	245	71	33	7	4	1	1041

Table 5.1 : (Contd.)

(i) Pallisamaj

length of preceding word in syllables	length of following word in syllables								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	41	91	39	6	3				180
2	91	198	108	19	10	1	1		428
3	41	101	59	15	5	1		1	223
4	6	16	10	3					35
5	2	9	6	1	1				19
6		2							2
7		1							1
8		1							1
total	181	419	222	44	19	2	1	1	889

(j) Pather Dabi

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	36	76	48	6				166
2	78	166	94	21	6			365
3	41	104	64	15	4		1	229
4	7	21	16			1		45
5	2	5	1					8
6		1						1
7		1						1
total	164	374	223	42	10	1	1	815

Table 5.1: (Contd.)

(k) Pather Panchali

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	38	84	35	13	1	171
2	85	230	101	29	8	453
3	31	103	66	12	6	218
4	10	25	19	1	2	57
5	3	10	6	2	2	23
total	167	452	227	57	19	922

(l) Devayan

length of preceding word in syllables	length of following word in syllables 1	2	3	4	5	6	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	39	107	31	11	1	1	190
2	116	290	80	29	7		522
3	22	91	25	9		1	148
4	9	26	12	5	5		57
5		8	5				13
6		1					1
total	186	523	153	54	13	2	931

Table 5.1 : (Contd.)

(m) Dristipat

length of preceding word in syllables	length of following word in syllables								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	26	64	28	10	4				132
2	65	147	89	34	7		2	1	345
3	33	82	46	24	4				189
4	12	39	18	6	4	1			80
5	2	7	6	6	2				23
6									
7		1	1						2
8	1								1
total	139	340	188	80	21	1	2	1	772

(n) Janantik

length of preceding word in syllables	length of following word in syllables						total
	1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	18	64	16	8	2	2	110
2	66	195	72	24	5	2	364
3	19	74	30	15	6		144
4	6	20	19	8			53
5	3	5	5		1		14
6		5					5
total	112	363	142	55	14	4	690

Table 5.1: (Contd.)

(o) Chacha-Kahini

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	22	70	28	8	2			130
2	81	250	80	18	7	1	1	438
3	16	97	33	11	1			158
4	5	22	10	1	1		1	40
5	2	5	3					10
6			1					1
7			1					1
total	126	444	156	38	11	1	2	778

(p) Deshe Videshe

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	32	78	26	8	2			146
2	79	248	80	26	7			440
3	19	81	29	11	2		1	143
4	11	26	7	5				49
5	4	4	2	1	1			12
6								
7					1			1
total	145	437	144	51	13		1	791

5.3.5. Consider the estimated proportion p_{mn} obtained from the (m,n) - cell of any subtable of, say Table 5.1. This is the estimate of the proportion of consecutive word-pairs in the text where the preceding word is m - syllabled and the following word is n - syllabled[1]. This estimate has the ratio form, for

$$p_{mn} = \frac{\sum_i n_{i(mn)}^{(o)}}{\sum_i n_i}$$

Here n_i is the number of words on the ith sample line and $n_{i(mn)}^{(o)}$ is the number of word-pairs formed from this ith line, with 0 words between them, where the preceding word is m-syllabled and the following word n-syllabled. Recalling the investigations of Chapter 2, Sections 2.4-2.5, it may be said that for the (m,n)-cells which have relatively high proportions, the estimates p_{mn} probably have the large sample properties of ratio estimates, even for the subsamples. But for the low-frequency cells of the two-way distributions, $n_{i(mn)}^{(o)}$ will have high c.v., for it may be 0 for most values of i, and 1 only occasionally, and c.v. of the sample average $\bar{n}_{i(mn)}^{(o)}$ may not be below 10%, as required. However, the biases may not be large, since the regressions of $n_{i(mn)}^{(o)}$ on n_i may hot be far from straight lines passing through the origin.

1/ Most of the conclusions are based on the autocorrelations estimated from two-way distributions, and the estimates p_{mn} have hardly been used by themselves.

5.3.6. As regards the autocorrelation coefficients based on such two-way tables, the situation is better, though not perfectly satisfactory. Denote the lengths of words on the i th sample line by $x_{i1}, x_{i2}, \ldots\ldots, x_{in_i}$, and those of the words paired with these (at gap s) by $y_{i1}, y_{i2}, \ldots\ldots, y_{in_i}$. Then consider the following estimates :

$$(1)\quad \frac{\sum_i \left(\sum_j x_{ij} y_{ij}\right)}{\sum_i n_i},$$

$$(2)\quad \frac{\sum_i \left(\sum_j x_{ij}^2\right)}{\sum_i n_i}, \qquad (3)\quad \frac{\sum_i \left(\sum_j y_{ij}^2\right)}{\sum_i n_i}$$

$$(4)\quad \frac{\sum_i \left(\sum_j x_{ij}\right)}{\sum_i n_i} \text{ and } (5)\quad \frac{\sum_i \left(\sum_j y_{ij}\right)}{\sum_i n_i}$$

Estimate (4) has been studied in Chapter 2 and it was found that the large sample properties of ratio estimates can be safely assumed even at the subsample level. Estimate (5) is clearly parallel[1/]. Even for the other three, it seems evident without any investigation, that the sample averages of numerator variables $\sum_j x_{ij} y_{ij}$, $\sum_j x_{ij}^2$ and $\sum_j y_{ij}^2$ would have c.v. below 0.1 at least for the combined samples. (These quantities are not so variable, measured by c.v., as $n_i^{(c)}$, say, of Chapter 4.) It may therefore be assumed with some confidence that all the five estimates given above have the large sample properties

---

1/ The observations y_{ij} ($i = 1,2,\ldots, 100$; $j = 1,2,\ldots, n_i$) form a probability sample linked with, and parallel to that given by the x_{ij}'s the latter have been used in the greater part of the present study.

(normality etc.)of ratio estimates, though not at the subsample level. Since all these estimates are consistent, it follows the estimated autocorrelation coefficient is at least a consistent estimate of the true value.

5.3.7. Now the numerator of the autocorrelation coefficient is the respective covariance. It will be assumed that the covariance estimate is normally distributed. Then, under the null hypothesis of zero autocorrelation, the estimated covariance is normally distributed with zero as mean. Hence the estimated autocorrelation coefficient would have zero as median, under the null hypothesis. But the sampling distribution may not be sufficiently near normal, for the present sizes of samples; also the bias may not be negligible as shown in the next Section.

5.3.8. This approximate result forms the basis of the tests of significance. For each work, the autocorrelations are available for four subsamples. Assuming this result for the subsamples — the conclusions have to be drawn with due reserve in view of the approximations introduced — one can call the autocorrelation coefficient significantly larger (smaller) than zero if all the subsample coefficients are positive (negative). (The level of significance is apparently 12.5%, but in reality 6.25%, since positive autocorrelations are expected.) In general, if the subsample coefficients are distributed with the true value as median, the range covered by them is a confidence internal for the true value with confidence coefficient 87.5%.

5.4.1. The Autocorrelation Coefficients for Bengali Prose : One may now examine the subsamplewise and combined estimates of r_1 presented in Table 5.2. These are based on the probability samples of words from the sixteen works listed.

Table 5.2: Estimates of autocorrelation coefficients (r_1) between lengths of consecutive words (in syllables), for sixteen works in Bengali prose*

works	no. of sample word-pairs	autocorrelation coeffcients r_1					simple average of sub-sample estimates	r_1 based on within line pairs (ss1-4)
		ss 1	ss 2	ss 3	ss 4	comb.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Shakuntala	696	0.031	0.038	0.004	-0.032	0.030	0.010	0.035
Sitar Vanavas	750	0.051	0.013	0.007	0.023	0.023	0.024	0.031
Durgeshnandini	577	0.152	0.052	-0.076	0.087	0.064	0.054	0.081
Visavriksha	611	0.184	0.085	0.179	0.200	0.139	0.162	0.096
Gora	889	0.018	0.081	0.060	0.077	0.061	0.059	0.066
Sheser Kavita	735	-0.030	0.060	0.037	-0.038	0.016	0.007	0.013
Char-Yari Katha	872	-0.007	-0.0005	0.062	0.044	0.028	0.025	0.023
Birbaler Halkhata	1041	0.085	0.067	0.105	-0.110	0.040	0.036	0.048
Pallisamaj	889	0.062	0.028	0.124	0.058	0.074	0.068	0.085
Pather Dabi	815	0.011	0.076	0.058	0.011	0.043	0.039	0.052
Pather Panchali	922	-0.038	0.076	0.093	0.193	0.088	0.081	0.085
Debayan	931	0.126	0.138	-0.019	0.083	0.087	0.082	0.103
Dristipat	772	-0.022	-0.006	0.274	0.020	0.070	0.067	0.058
Janantik	690	-0.026	0.085	0.018	0.170	0.076	0.062	0.103
Chacha Kahini	778	-0.060	0.132	-0.055	0.121	0.048	0.035	0.075
Deshe Videshe	791	0.107	0.012	-0.010	0.120	0.063	0.057	0.077
Average						0.0594	0.0542	0.0645

* based on probability samples of words falling on 100 randomly selected lines from each work.

5.4.2. There seems to be a small negative bias in the estimates. For in only 2 out of 16 cases, the simple average of the subsample estimates ⌈col.(8)⌋ exceeds the combined estimate ⌈col.(7)⌋. The

probability of not more than 2 heads in 16 tosses of an unbiased coin is only about 0.002.

5.4.3. The straight average of the combined estimates of all sixteen works is 0.0594, and the corresponding average of the simple averages of subsample estimates is 0.0542. This means a difference of a little below 10%. Assuming that the biases in the estimates are of order 1/n, n being the sample size, an unbiased estimate of the average of the true r_1's of all sixteen works seems to be

$$0.0594 + \frac{1}{3}\left[0.0594 - 0.0542\right] = 0.0611$$

[vide Section 2.5 of Chapter 2]. Such corrections for bias can be made for each work, but the estimates would be less reliable. In any case, the bias does not seem to be really important, being much smaller than sampling errors.

5.4.4. If the presence of bias be neglected, and if one takes an overall view covering all sixteen works, the value of r_1 is at once seen to be significantly positive. For all the sixteen combined estimates of r_1 are positive, and so are all the averages of subsample estimates. Probability of 16 heads in 16 tosses of an unbiased coin is extremely small (1/65536). Again, out of the 64 subsamplewise estimates in the whole table, only 14 have the minus sign. The probability of 50 heads or more in 64 tosses of an unbiased coin is, using the normal approximation, the area above the abscissa 4.375 of the N(0,1) distribution, that is, $0.0^5 6$. That the estimates have a small negative bias actually strengthens these conclusions regarding significance.

5.4.5. The figures for individual works are much less conclusive. Assume that the subsample r_1's are distributed with zero as median under the null hypothesis of zero autocorrelation. The probability of three r_1's being positive is $\frac{1}{8}$. So only if all the subsample estimates are positive, one can talk about significance of the estimated r_1 (at the 6.25% level of significance). In view of the negative bias demonstrated above, this test would probably err on the safe side. Such significance is seen for 6 out of the 16 works, viz., 'Sitar Vanavasa', 'Visavriksha', 'Gora', 'Pallisamaj', 'Pather Dabi' and 'Deshe Videshe'.

5.4.6. The test would somewhat improve in power if all the 8 subsample estimates relating to the same author could be considered together. This is a dubious procedure, and certainly not meaningful if the works are not comparable in subjectmatter or style. Anyway, this approach gives significant results at 5% level if at least 7 out of the 8 signs are positive, using the same assumptions regarding the sampling distribution; but if all 8 estimates are positive, the significance reaches the 1% level. It is found that, taken together, the 8 estimates show significantly positive r_1 only for Vidyasagar (7 positive), Bamkimchandra (7 positive) and Saratchandra (8 positive), but not for any other author. It may also be argued that the two works by Bamkimchandra are written in appreciably different languages.

5.4.7. It is very difficult to say anything about the significance of between work or between author differences in r_1.

5.4.8. Finally, one may notice the "within line" coefficients presented, for the sake of interest, in col.(9)of Table 5.2. These are the values of r_1 based on all pairs of consecutive words underlying figures in col.(7) excepting the 100 pairs formed by the last words of sample lines with the first words of the following lines. So sample sizes for these coefficients are 100 less than the sizes shown in col. (2). These estimates are seen to be fairly close to the combined estimates presented in col. (7). The "within line" coefficients tend to be slightly higher than the combined coefficients, but the differences are small for practical purposes.

5.4.9. One may now examine the estimates of autocorrelation coefficients of higher orders, r_2, r_3, r_4 and r_7, presented for four selected works in Table 5.7. The estimates are given by subsamples and also for the combined sample. The coefficients r_1 are reproduced from Table 5.2. The simple averages of subsample estimates are also given in col. (9), Col. (10) shows the within line coefficients, that is, coefficients based on all word-pairs with gap 0, 1, 2, 3 or 6 which could be formed from words on the sample lines, that is, without using the words on the following lines. For the coefficient r_j (j= 1, 2, 3, 4, 7) the sample size is approximately that shown in col. (2) less j times 100.

Table 5.3: Joint distributions of lengths of neighbouring words at gap one, based on probability samples of words from four works in Bengali prose.

(a) Visavriksha

length of the preceding word (syllables)	length of the second next word (syllables) 1	2	3	4	5	6	7	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	20	40	30	6	3			99
2	37	99	73	28	7	3	1	248
3	32	66	51	22	4		1	176
4	7	21	20	9	3	1	1	62
5	6	3	8	3				20
6	2	2						4
7			2					2
total	104	231	184	68	17	4	3	611

(b) Gora

length of the preceding word (syllables)	length of the second next word (syllables) 1	2	3	4	5	6	7	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	23	71	34	7	2	1		138
2	66	217	106	25	9		1	424
3	28	112	79	17	10	1		247
4	7	27	19	3	2	1		59
5	6	6	8					20
6			1					1
total	130	433	247	52	23	3	1	889

Table 5.3: (Contd.)

(c) Pather Dabi

length of the preceding word (syllables)	length of the second next word (syllables)							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	41	61	53	10	1			166
2	71	175	97	19	4			366
3	42	106	61	11	7	1		228
4	3	25	14	2			1	45
5	1	3	4					8
6				1				1
7	1							1
total	159	370	229	43	12	1	1	815

(d) Dristipat

length of the preceding word (syllables)	length of the second next word (syllables)								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	21	75	23	13	1				133
2	71	149	84	32	6	1	1	1	345
3	30	87	45	19	5	2			188
4	14	33	17	10	6				80
5	3	8	5	6	1				23
6									0
7			2						2
8		1							1
total	139	353	176	80	19	3	1	1	772

Table 5.4: Joint distributions of lengths of neighbouring words at gap two, based on probability samples of words from four works in Bengali prose.

(a) Visavriksha

length of the preceding word (syllables)	length of the third next word (syllables)								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	14	37	32	14	1	1			99
2	40	92	81	25	8	1		1	248
3	29	69	53	18	5	1	1		176
4	14	22	18	6		1	1		62
5	4	7	4	4	1				20
6	2	1		1					4
7	1		1						2
total	104	228	189	68	15	4	2	1	611

(b) Gora

length of the preceding word (syllables)	length of the third next word (syllables)							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	25	70	34	7	1	1		138
2	59	215	114	25	9	2		424
3	32	120	69	17	8		1	247
4	6	29	18	4	2			59
5	3	6	10	1				20
6			1					1
total	125	440	246	54	20	3	1	889

Table 5.4: (Contd.)

(c) Pather Dabi

length of the preceding word (syllables)	length of the third next word (syllables)							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	25	77	51	10	2	1		166
2	83	164	100	14	5			366
3	40	101	64	16	5	1	1	228
4	8	17	18	2				45
5	1	3	4					8
6		1						1
7	1							1
total	158	363	237	42	12	2	1	815

(d) Dristipat

length of the preceding word (syllables)	length of the third next word (syllables)								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	32	55	24	18	3	1			133
2	59	161	82	32	8	2		1	345
3	37	79	48	16	7	1			188
4	11	39	19	7	3		1		80
5	4	8	7	4					23
6									0
7		2							2
8	1								1
total	144	344	180	77	21	4	1	1	772

Table 5.5: Joint distributions of lengths of neighbouring words at gap three, based on probability samples of words from four works in Bengali prose.

(a) Visavriksha

length of the preceding word (syllables)	length of the fourth next word (syllables)								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	17	37	27	12	6				99
2	40	95	76	29	3	4	1		148
3	32	64	52	24	2			2	176
4	8	23	21	8	2				62
5	4	8	6	2					20
6		2	1	1					4
7			2						2
total	101	229	185	76	13	4	1	2	611

(b) Gora

length of the preceding word (syllables)	length of the fourth next word (syllables)							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	20	67	43	4	2	2		138
2	65	216	110	26	5	1	1	424
3	28	121	74	16	8			247
4	11	30	13	4	1			59
5	2	11	3	3	1			20
6		1						1
total	126	446	243	53	17	3	1	889

Table 5.5: (Contd.)

(c) Pather Dabi

length of the preceding word (syllables)	length of the fourth next word (syllables)						total
	1	2	3	4	5	6	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	42	69	46	8	1		166
2	55	161	117	24	8		365
3	52	93	68	10	4	2	229
4	5	23	13	3	1		45
5	1	5	2				8
6		1					1
7	1						1
total	156	352	246	45	14	2	815

(d) Dristipat

length of the preceding word (syllables)	length of the fourth next word (syllables)								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	23	63	30	11	5	1			133
2	71	155	81	27	9	2			345
3	33	78	42	27	4	2	1	1	188
4	11	35	20	11	3				80
5	4	11	4	3	1				23
6									0
7	1	1							2
8		1							1
total	143	344	177	79	22	5	1	1	772

Table 5.6: Joint distributions of lengths of neighbouring words at gap six, based on probability samples of words from four works in Bengali prose.

(a) Visavriksha

length of the preceding word (syllables)	length of the seventh next word (syllables) 1	2	3	4	5	6	7	8	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	19	39	31	7	2			1	99
2	47	92	74	29	3	1		2	248
3	32	71	52	19	1	1			176
4	4	23	20	8	5	1	1		62
5	1	8	4	6	1				20
6		3		1					4
7		1			1				2
total	103	237	181	70	13	3	1	3	611

(b) Gora

length of the preceding word (syllables)	length of the seventh next word (syllables) 1	2	3	4	5	6	7	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	15	81	34	7	1			138
2	65	209	113	29	4	3	1	424
3	36	105	83	19	4			247
4	7	31	19	1	1			59
5	3	9	6	1	1			20
6				1				1
total	126	435	255	58	11	3	1	889

Table 5.6: (Contd.)

(c) Pather Dabi

length of the preceding word (syllables)	length of the seventh next word (syllables) 1	2	3	4	5	6	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	29	72	51	9	4	1	166
2	83	151	110	18	4		366
3	42	105	69	9	2	1	228
4	8	18	11	7	1		45
5	1	2	2	2	1		8
6	1						1
7				1			1
total	164	348	243	46	12	2	815

(d) Dristipat

length of the preceding word (syllables)	length of the seventh next word (syllables) 1	2	3	4	5	6	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	25	70	21	11	5	1	133
2	66	145	82	38	12	2	345
3	26	89	46	18	7	2	188
4	13	33	16	12	4	2	80
5	5	11	4	3			23
6							0
7				2			2
8		1					1
total	135	349	169	84	28	7	772

Table 5.1 Estimates of autocorrelation coefficients r_1, r_2, r_3, r_4 and r_7 between lengths of neighbouring words (in syllables) for four works in Bengali prose*

work	no. of sample word-pairs	coefficient	autocorrelation coefficients ss 1	ss 2	ss 3	ss 4	comb.	simple average of sub-sample estimates	coefficient based on within line pairs (ss1-4)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Visavriksha	611	r_1	0.184	0.085	0.179	0.200	0.139	0.162	0.096
		r_2	0.122	0.068	0.022	-0.022	0.047	0.047	0.080
		r_3	0.126	-0.092	-0.111	-0.154	-0.069	-0.058	-0.024
		r_4	-0.083	0.059	0.010	-0.048	-0.0002	-0.015	-0.007
		r_7	0.122	0.170	0.081	0.064	0.112	0.109	0.177
Gora	889	r_1	0.018	0.081	0.060	0.077	0.061	0.059	0.066
		r_2	0.083	0.061	0.046	0.086	0.068	0.069	0.072
		r_3	0.062	0.100	0.070	0.046	0.071	0.070	0.038
		r_4	-0.042	0.107	0.118	-0.124	0.023	0.015	0.056
		r_7	0.121	0.059	0.030	-0.032	0.042	0.044	0.069
Pather Dabi	815	r_1	0.011	0.076	0.058	0.011	0.043	0.039	0.052
		r_2	0.102	0.282	-0.078	0.121	0.057	0.107	0.066
		r_3	0.091	-0.062	-0.013	0.024	0.012	0.010	-0.024
		r_4	-0.084	-0.064	0.088	0.164	0.019	0.026	0.053
		r_7	0.048	0.132	0.057	0.001	0.031	0.060	0.087
Dristipat	772	r_1	-0.022	-0.006	0.274	0.020	0.070	0.067	0.058
		r_2	-0.005	0.121	0.187	0.073	0.098	0.094	0.073
		r_3	0.033	0.001	0.137	0.104	0.072	0.069	-0.017
		r_4	0.227	-0.080	0.168	0.001	0.076	0.079	-0.025
		r_7	0.087	0.115	0.078	-0.004	0.065	0.069	-0.028

* based on probability samples of words falling on 100 randomly selected lines from each work.

5.4.10. It is not safe to reach many definite conclusions from the estimates in Table 5.7. So comments will be made very briefly. One point is evident : even r_7 is statistically significant. The combined estimates are positive for all four works and among the sixteen sub-sample estimates only two are negative. The average value of r_7 seems to be about 0.06, only a little below the average value of r_1 for these four works (0.08). Actually, the correlation coefficients r_1, r_2 etc. do not show any marked tendency of the correlogram rapidly falling to zero. For "Dristipat" the coefficients hardly show any trend, and the same holds for "Pather Dabi" as well; "Gora" seems to show a decreasing trend, but "Visavriksha" suggests a clear cycle, with r_3, r_4 both negative but $r_7 > 0$. If the estimates for "Visavriksha" are taken very seriously, r_3 and r_4 do not come out significant and the possibility of oscillations in the correlogram has to be recognised. In any case, it seems that the autocorrelations do not fall very rapidly to zero but persist for intervals of a few words. This appears plausible in view of the patchy nature of works discussed in the following sections; since the patches generally contain much more than 6 words, two words at gap 6 are often both in the same patch.

5.5.1. Evidence from short Bengali Passages : Two passages were selected from "Visavriksha" and five from "Dristipat". All the passages were short, and comprised 200 to 250 words each. The selection of passages was purposive. Tables 5.8, 5.9 and 5.10 are based on these passages. The nature of the passages can be seen from Table 5.8

particularly col.(3). Most of the passages are indeed "patches" and were found in a preliminary search for different types of patches.

5.5.2. The first objective was to see the autocorrelations within homogeneous passages of moderate length. The distinction between such autocorrelations and the autocorrelations for entire works was drawn in para 5.3.1 above. The autocorrelations for entire works have been estimated from probability samples of words; the sampling theory has been found to be a little complicated (vide Sections 5.3 and 5.4). No such difficulty arises with continuous passages, but the autocorrelations become less meaningful.

5.5.3. The other objective was to examine the presence of "patches" with unusually high or unusually low levels of word-length. The question has been discussed in para 5.2.5, and it has been pointed out that runs of conversational words and runs of other words tend to form such patches in fiction (vide Chapter 4, Section 4.1 for details).

5.5.4. The second objective may be taken up first. It is remarkable that a preliminary search can bring to light such unusual patches from either work. Consider the two patches from 'Visavriksha'. The averages of word-length are 1.99 and 3.08 syllables, while the average for the whole work is nearly 2.47 [Table 3.1 of Chapter 3]. The differences are striking[1].

1/ And so is the second passage. For no work/text covered in Chap.3 gave an average above 2.9, except one poem, "Varsamangal", where the average is 3.35 (vide Table 3.3). So far as prose works are concerned, the highest average is that for "Shakuntala" viz., 2.704 (vide Table 3.1). Again, the longest word met with so far is of 14 syllables, in the poem, "Meghadut" (vide Table 3.4). Here, the passage of 241 words shows 7 words above 10 syllables, the highest being 18-syllables [vide Table 5.9(a)].

Table 5.8: Word-length distributions in short passages selected from "Visavriksha" and "Dristipat"

work	passage no.	type of passage	no.of words	percentage of words by length in syllables						average length (syllables)	s.d. (syllables)
				1	2	3	4	5	6		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Visavriksha	1	short words, soliloquy	257	29.57	46.30	21.01	1.95	1.17		1.988	0.830
	2	very long words, non-conversation	241	13.70	39.43	23.25	13.29	2.90	7.43*	3.083	1.468
Dristipat	1	short words, almost wholly conversation	216	28.24	51.85	16.67	3.24			1.949	0.759
	2	short words, wholly non-conversation	218	19.27	50.45	18.81	8.26	2.75	0.46	2.262	0.982
	3	long words, non-conversation	212	13.68	31.13	31.61	17.92	4.72	0.94	2.717	1.110
	4	medium size words, non-conversation	211	13.74	46.46	24.64	12.32	2.37	0.47	2.446	0.984
	5	medium size words, mixed	221	15.38	44.80	27.15	10.86	1.81		2.389	0.933

* Includes many words with more than 6 syllables [vide Table 5.9(a)]

Table 5.9: Joint distributions of lengths of consecutive words for two of the short passages listed in Table 5.8

(a) Visavriksha (passage no.2)

length of preceding word (syllables)	length of following word in syllables															total	average length of following words
	1	2	3	4	5	6	8	9	10	13	14	15	16	17	18		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
1	3	6	7	11			2	1			2			1		33	4.55
2	12	42	25	10	3	1		1				1				95	2.69
3	11	25	11*	6	1	1	1									55	2.45
4	3	15	8	3	1	2										32	2.69
5	2	2	3													7	2.14
6		2	2		1											5	3.00
8		1			1										1	3	8.33
9						1			1							2	8.00
10													1			1	16.00
13		1														1	2.00
14	2															2	1.00
15				1												1	4.00
16										1						1	13.00
17		1														1	2.00
18				1												1	4.00
total	33	95	56	32	7	5	3	2	1	1	2	1	1	1	1	241	3.08

* Includes the circular pair

Table 5.9: (Contd.)

(b) Dristipat (passage no.5)

length of preceding word (syllables)	length of following word (syllables)					total	average length of following word
	1	2	3	4	5		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	6	15	9	4*		34	2.32
2	17	51	24	6	1	99	2.22
3	9	19	20	9	3	60	2.63
4	2	11	7	4		24	2.54
5		3		1		4	2.50
total	34	99	60	24	4	221	2.39

* Includes the circular pair

Table 5.10: Wald-Wolfowitz test for circular* autocorrelation coefficients between lengths of neighbouring words in terms of syllables, estimated from passages listed in Table 5.8.

work	passage no.	no. of words	autocorrelation coefficient r_1			autocorrelation coefficient r_{10}	
			r_1	standard error	critical ratio	r_{10}	critical ratio
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Visavriksha	1	257	-0.023 (-0.115)	0.064	-0.295	-0.012 (0.012)	-0.118
	2	241	0.139 (0.139)	0.062	2.305	-0.054 (-0.055)	-0.868
Dristipat	1	216	-0.085	0.068	-1.185	-	-
	2	218	0.081	0.067	1.279	-	-
	3	212	0.007	0.065	0.074	-	-
	4	211	0.088	0.068	1.297	-	-
	5	221	0.133	0.067	1.982	-	-

* The figures within brackets are the noncircular coefficients

5.5.5. Can it be that the word-length series is really random and that such differences are entirely due to chance ? [Vide Kendall and Babington Smith (1939), Introduction, for local patches in random sampling number tables.] The following rough test nearly dismisses this possibility. At a rough estimate, Visavriksha has about 35000 words in all. It can therefore be spilt up into 145 mutually exclusive passages of about 240 words each. The overall average for "Visavriksha" is about 2.46 and the overall s.d. 1.07 (vide Table 3.1). If the series were completely random, the average word-length of these 240-word passages would be distributed normally, by the Cemtral Limit Theorem, with mean 2.46 and s.d. $\frac{1.07}{\sqrt{240}}$, that is, 0.069. It is extremely unlikely that among the 145 means, there would be one (3.08) at 9 σ above the overall mean and another (1.99) at about 7 σ below it[1]. And probably passage no.1 is not the one with the lowest average in "Visavriksha".

5.5.6. Consider now the passages numbered 1 and 3 from "Dristipat". There are about 185 mutually exclusive passages of 220 words each in this work of about 40500 words. Under the null hypothesis of perfect randomness of the word-length series, the averages of these passages would be normally distributed with mean approximately 2.40 and s.d. $\frac{1.04}{\sqrt{220}}$ = 0.07. Again, it is extremely improbable that among 185 passages, there would be one with mean 2.72, that is, over 4 σ above

1/ If all sets of 240 consecutive words are considered instead of the particular set of mutually exclusive passages, the probability of such extreme types of passages would be somewhat increased, but this refinement cannot alter the conclusion.

the overall mean, and another with mean 1.95, which means more than 6 σ below it. And in this case, probably, the selection can be "improved" in both the directions.

5.5.7. The obvious explanation is that the word-length series is not fully random. Due to differences between conversational and other words and due to other reasons like variation of topic, more or less conspicuous patches are there with medium or (unusually) high or (unusually) low averages of word-length. To put the same thing in another way, the true s.d.'s of the passage means discussed in paras 5.5.5-6 are higher than 0.069 or 0.07 due to intra-passage correlations. The multiplying factor is $\sqrt{1 + (n-1)\bar{\rho}}$, where n is the passage length in words and $\bar{\rho}$ the average intra-passage correlation between lengths of all pairs of distinct words. It may be that $\bar{\rho}$ is nearly 0.02 (vide para 5.2.4.). The multiplying factor would then be about 2.5 for passages of 250 words. The deviations of 7σ or 9σ from the mean discussed in paras 5.5.5-6 would really mean deviations by about 3σ, which are relatively frequent.

5.5.8. Finally, one should also notice the averages for the other passages from "Dristipat". Apparently patches can be found at will having $\bar{x}$ anywhere in the range 2 to 2.7.

5.5.9. As regards the first objective stated in para 5.5.2, the relevant figures are presented in Table 5.10 and some joint distributions are shown in Table 5.9. The coefficient r_1 is negative for one of the two "Visavriksha" passages and also for one of the five

"Dristipat" passages. So the coefficient cannot be considered as significantly greater than zero merely by noting the signs. [The conditional expectation of r_1 is $-\frac{1}{n-1}$, and not zero, but this makes very little difference.]

5.5.10. The average r_1 for the two "Visavriksha" passages is about 0.06, and the average for the five "Dristipat" passages about 0.04. These may be, to a first approximation, the true values of the within passage autocorrelations[1/]. But it is doubtful whether they are significantly above zero. We have applied the Wald-Wolfowitz (1943) test for judging the significance of autocorrelation coefficients. [Vide Chapter 9, Section 9.2, where this test is described along with several others.] One out of the two critical ratios for "Visavriksha" exceeds the (two-sided) 5% level, but the other is negative; the sum of the two critical ratios is 2.010 which is distributed as $N(0, \sqrt{2})$ and hence not significant at the one-sided 5% level. Only one out of the five 'Dristipat' passages (no.5) shows a significantly positive r_1. The sum of the five critical ratios is 3.447; since this should be distributed as $N(0, \sqrt{5})$, the sum is not at all significant. But if the 7 critical ratios are considered together, one gets a significant result. The sum of the seven ratios is 5.457, and since this should

---

1/ These are less than the corresponding estimates for entire works given in Table 5.2, viz., 0.139 for "Visavriksha" and 0.070 for "Dristipat". The differences are not significant, but they are in the expected direction.

be a N(0, $\sqrt{7}$)-variate under the null hypothesis of randomness, the sum would come out significant at the two-sided 5% level.

5.5.11. The value of r_{10} given for two "Visavriksha" passages are both negative, but they are nonsignificant and near zero. This appears plausible.

5.6.1. Some Investigation on English: A probability sample of words was drawn from Chapters 1 to 32 (pp.1 - 199) of the novel, "Pride and Prejudice", by Jane Austen; 200 lines were selected at random and words falling on selected lines comprised the probability sample of words. This was split into four independent and interpenetrating sub-samples, each based on 50 randomly selected lines. A systematic sample of words was also drawn from the same population; this comprised words falling on the 3rd line (from top) of every page, starting from page 1 and ending in page 199. A systematic sample was also drawn from "The Tale of Two Cities" by Charles Dickens : This comprised words falling on the 5th line from top of every 5th page, beginning with page 5. This sample was split into two subsamples by assigning the lines on pages 5, 15, 25, ... to subsample 1, and the remaining lines to subsample 2.

5.6.2. Appendix 3 presents the word-length distributions obtained from these samples and subsamples; shows the agreement between probability and systematic samples from "Pride and Prejudice"; and brings out the difference between conversational words and other

words in this English novel. The same material will be utilised here for studying the autocorrelation coefficient r_1 in English, measuring word-length in letters. /For "Pride and Prejudice", only the probability sample will be used./ In addition to these, some short continuous passages were studied, one from "Pride and Prejudice" and two from Shakespeare's "Othello".

5.6.3. The joint distributions of lengths of consecutive words are presented in Tables 5.11 and 5.12 for "Pride and Prejudice" (Chap. 1-32) and "The Tale of Two Cities" respectively. The estimated autocorrelation coefficients r_1 are shown in Table 5.13. The estimates of r_1 are given for the combined samples as well as for the two or four subsamples; but the joint distributions are shown only for the combined sample. For reasons which will become clear below, the joint distribution for "Pride and Prejudice" given in Table 5.11(a) was split up into three joint distributions, of which two are shown as Tables 5.11(b) and (c); the table (b) covers all word-pairs where both words are conversational, and the table (c), all word-pairs where both the words are not conversational. The joint distribution of the small number of "mixed" pairs is not shown. The corresponding estimates r_1 are also shown in Table 5.13 separately by subsamples and for the combined sample.

Table 5.11 Joint distributions of lengths of consecutive words in "Pride and Prejudice" (Chap. 1-32) based on the probability sample of words occurring on 200 randomly selected lines.

(a) all word-pairs

length of preceding word (letters)	length of following word in letters														total	average length of the following word
	1	2	3	4	5	6	7	8	9	10	11	12	13	14		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1		2	4	20	6	8	5	1	2	2		2			52	5.37
2	11	38	80	63	44	21	27	18	16	6	3	2	3		332	4.69
3	13	56	69	86	40	46	40	26	22	11	2	4		1	416	4.88
4	12	70	89	45	23	21	18	17	14	6	2	4	1		322	4.26
5	4	44	33	23	7	13	12	7	9	2	2		2		158	4.34
6	1	35	43	20	13	5	5	3	4			2			131	3.77
7	6	26	36	19	11	5	5	3	6	2	1	1			121	3.99
8	6	20	25	14	7	5	2	3		2					84	3.57
9	3	26	20	14	7	7	4	5			1				87	3.76
10		6	7	7	2	2	4	3	1					1	33	4.79
11		4	1	1		3		3							12	4.75
12	1	9	2	2	1	1	1		1						18	3.33
13		2	1	1											4	2.75
14		1	1												2	2.50
total	57	339	411	315	161	137	123	89	75	31	11	15	6	2	1772	4.41

Table 5.11: (Contd.)

(b) all pairs of conversational words

length of preceding word (letters)	length of following word in letters													total
	1	2	3	4	5	6	7	8	9	10	11	12	13	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1		2	3	16	3	5	3			1		1		34
2	5	22	37	43	22	11	13	8	8	3	3			175
3	10	38	26	46	17	21	11	13	7	3		2		194
4	10	38	44	17	13	9	9	6	6	3	1		1	157
5	3	22	18	8	2	4	4	1	2	1			1	66
6		18	16	13	5	4	1	1	1			1		60
7	3	12	18	3	1	2	2	2	1	1				45
8	3	8	11	2	4	3	1			1				33
9	2	9	3	9		3	1				1			28
10		1	3	3		2	2							11
11		1	1	1		1		2						6
12	1	4	1			1								7
13		1												1
total	37	176	181	161	67	66	47	33	25	13	5	4	2	817

Table 5.11: (Contd.)

(c) all pairs of non-conversational words

length of preceding word (letters)	length of following word in letters														total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1			1	4	3	3	2	1	2	1		1			18
2	6	16	43	19	22	10	14	10	8	3		2	3		155
3	3	18	43	39	23	25	29	13	15	8	2	2		1	221
4	2	32	43	27	10	12	9	11	8	3	1	4			158
5	1	22	15	15	5	9	8	6	6	1	2		1		91
6	1	17	25	7	8	1	3	2	3			1			67
7	2	14	18	15	9	3	3	1	5	1	1	1			73
8	2	12	14	12	3	2	1	3		1					50
9	1	15	17	4	6	4	2	5							53
10		4	4	3	2		1	3	1					1	19
11		3				2		1							6
12		5	1	2	1		1		1						11
13		1	1	1											3
14		1	1												2
total	18	160	226	148	92	71	73	56	49	18	6	11	4	2	934

Joint distribution of lengths of consecutive words in "A Tale of Two Cities" based on a systematic sample of words (occurring on 76 lines).

length of preceding word (letters)	length of following word in letters														total	average length of following word
	1	2	3	4	5	6	7	8	9	10	11	12	13	14		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1			1	9	3	4				1					18	4.89
2	6	25	34	26	14	10	6	4	3	1	1		1		131	4.02
3	5	16	26	33	21	18	13	11	7	3	3		2		158	5.01
4	6	28	35	35	14	11	7	5	4	2				1	148	4.07
5	1	16	19	11	12	6	2	2	3	1					73	4.07
6		15	14	14	4	3	5	1	1		1	1			59	4.08
7		5	18	4	2	4	1	1				1			36	3.92
8		12	5	4	2	1				1					25	3.24
9		4	3	4	3	2	1	3	1		1				22	5.05
10		5		1	1		1								8	3.25
11		4	1												5	2.20
12			1	2											3	3.67
13		2													2	2.00
14		1													1	2.00
total	18	133	157	143	76	59	36	27	19	9	6	2	3	1	689	4.26

15. Estimation of autocorrelation coefficients r_1 between lengths (in letters) of consecutive words in two English works.

work	type of sample	sub-sample	no. of sample lines	no. of word-pairs			estimate of r_1		
				all pairs	both words conv.	both words non-conv.	all pairs	both words conv.	both words non-conv.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Pride and Prejudice (Chaps.1-32)	proba-bility	1	50	460	198	256	-0.125	-0.133	-0.150
		2	50	412	198	210	-0.160	-0.152	-0.185
		3	50	440	154	282	-0.198	-0.195	-0.219
		4	50	460	267	186	-0.132	-0.156	-0.144
		comb.	200	1772	817	934	-0.151	-0.150	-0.174
A Tale of Two Cities	syste-matic	1	38	337	-	-	-0.131	-	-
		2	38	352	-	-	-0.092	-	-
		comb.	76	689	-	-	-0.110	-	-

5.6.4. As in the study on Bengali prose reported earlier in this Chapter, the pair formed by the last word of each sample line and the first word of the next line in the text was included in the sample of word-pairs, so as to give proper representation to all word-pairs in the sampled text (vide Section 5.3 supra).

5.6.5. We first consider the estimates of r_1 based on all word-pairs. The subsamplewise and combined estimates are all negative and evidently the value of r_1 is significantly lower than zero. The deviation from zero is also quite appreciable, the estimates being of the order of -0.1 and -0.15. The negative correlation is also apparent from the declining trend in the averages presented in the marginal columns of Tables 5.11(a) and 5.12.

5.6.6. This result partly contradicts Fucks (1954) who showed that the value of r_1 is practically zero in English and German. But Fucks considered only one English work, viz., Shakespeare's Othello, which is a drama and partly in verse. /The estimate of r_1 was -0.02992./ Fucks measured word-length in syllables whereas the present author measured word-length in letters, but this distinction may not be really important.

5.6.7. Two short continuous passages were selected for examination from Othello. The first contained the first 224 speech words of the drama and was entirely in verse; the second was from scene 5.2, line 23 onwards, where Othello and Desdemona are talking before the

killing takes place. The estimates of (circular) r_1 and the results of applying the Wald-Wolfowitz non-parametric test are shown in Table 5.14. This table also covers one short passage selected from "Pride and Prejudice". The mean word-length is 4.44 letters for this last-mentioned passage, which is more or less the average for English prose [vide Chapter 1, para 1.2.14]. The averages for the Othello passages are appreciably lower; it is wellknown (vide Elderton, 1949) that word-length in Shakespeare is smaller, on the average, than in most other writers in English.

Table 5.14: Wald-Wolfowitz tests of circular autocorrelation coefficients r_1 between lengths of consecutive words in terms of letters, estimated from three short passages in English.

work	passage no.	no. of words	circular r_1	standard error	critical ratio	mean word-length (letters)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Othello	1	224	-0.008	0.107	-0.033	4.11
	2	276	0.064	0.060	1.128	3.83
Pride and Prejudice	1	211	-0.159	0.073	-2.105	4.44

5.6.8. It is apparent that the value of r_1 is significantly negative for the passage from "Pride and Prejudice"; it is also nearly equal to that for the entire work (shown in Table 5.13). But for the passages from Othello, the values of r_1 are small in absolute terms and non-significant. The average of two values is about 0.03. Fucks'

estimate of r_1 for the entire work was -0.02992. In view of the sampling errors of the estimates presented here, the difference does not appear to be really wide.

5.6.9. It appears that r_1 may be -0.1 or -0.15 for most English prose works, and the figure for Othello may not be really typical.

5.6.10. The impression that the series of word-lengths is very nearly random seems to have gained ground (Fucks, 1954). This may not be correct for most works in English prose, as shown in the foregoing paras. We may also point to the distinction between conversational and other words in "Pride and Prejudice" demonstrated in Appendix 3.

5.6.11. The distinction between conversational and other words in English should give rise to positive autocorrelation coefficients r_1, r_2 etc. It is evident that the two classes of words occur in long runs. Among the 1772 pairs of consecutive words obtained from 200 randomly selected lines from "Pride and Prejudice", there were only 21 pairs where one word was conversational and the other not [vide Table 5.13]. It could therefore be expected that the overall autocorrelation coefficient r_1 would be positive even though it may be nearer zero within conversational or within non-conversational matter [vide para 5.2.5]. These expectations are not fulfilled. Consider the estimates for "Pride and Prejudice" presented in Table 5.13. The overall r_1 is appreciably and significantly negative. Also, the

values of r_1 seem to be very much the same within conversational matter and also within other matter as for the overall text. This line of study seems to merit further investigation.

5.6.12. It was suspected at first that this negative autocorrelation is primarily due to very long words being mostly preceded and/or followed by very short grammar words. But this impression was not quite correct. Thus, excluding the 14 pairs from the two-way distribution for the "Pride and Prejudice" passage [Table 5.11(a)] where at least one word had length 13 or more, the value of r_1 came to about -0.131, which is not much smaller than the original value -0.159.

5.6.13. The real explanation of the negative r_1 seems to be the presence in English of two distinct streams of words, viz., grammar words, which are shorter and content words which are longer, on the average, and the tendency of words of these two classes to occur alternately [Miller, Newman and Friedman (1958), Herdan(1956, pp.111-5)]. Compared with this, the presence of alternate patches of shortish (conversational) and longish (other) words, seem to have much smaller effect.

5.6.14. The value of r_1 for Bengali prose, it may be recalled, is of the order of 0.06. The difference between English and Bengali cannot be due to the difference between syllables and letters — word-length has been measured in letters for English and in syllables for Bengali. The difference must be largely due to English having

much fewer inflexions and hence much more of short grammar words which tend to alternate with the longer content words. This feature is not at all conspicuous in Bengali, where the presence of alternate patches of shortish and longish words seems to be relatively quite important.

Chapter 6 : The Form of the Word-length Distribution

6.1.1. Introduction: Chapter 3 presents the word-length distributions for many works or texts in Bengali prose. The proportions p_x of words of length x (x= 1, 2, ...) were estimated from nearly random samples of words. Word-length x was measured in syllables. It is the purpose of the present Chapter to study the functional form of these word-length distributions, that is to say, how far theoretical distributions, like the Poisson, can fit these observed distributions for Bengali prose[1/].

6.1.2. Elderton (1949) reports that the geometric-distribution $p_x = ab^x$ (Feller, 1957, Secn. VI.8, XVII.6) fits certain word-length distributions like those from Fitzgerald's "Rubaiyat" of Omar Khayyam, xxxxxxxxx but is generally useless for word-length distributions for English texts.

6.1.3. A comprehensive attempt was made by Fucks (1955), who concluded that if x denotes the length of a word in syllables, then x -1 is approximately distributed as a Poisson variate for eight out of the nine languages examined by him, viz., English, German, Esperanto, Greek, Japanese, Russian, Latin and Turkish, Arabic proving to be the only exception. Fucks also proposed a stochastic model, leading to

---

1/ For two works in Bengali prose, word-length was also measured by the number of letters. The findings are reported in Appendix 1 ; among other things, the form of the distributions is also examined there.

this Poisson law, for allocating the total number of syllables among a total number of words, at least one syllable being allocated to each word; and a Galton-Board-like apparatus was used to illustrate the model. More general and more complicated models were put forward by Fucks in pp.163- 7 of the same paper.

6.1.4. The lognormal distribution was found somewhat suitable for certain distributions for English by Williams (1956) and also by Herdan (1958)[1]. In both cases, x was the number of letters comprising the word.

6.1.5. Since p_2 is considerably larger than p_1, the geometric law fails completely for the data on Bengali. The Poisson law also gave a poor fit, in general, excepting for one or two works; this will be reported in detail in Section 6.3. The goodness of fit criteria are explained in Section 6.2. Examination of Fucks' data shows that the Poisson law gives only a crude first approximation for most of the nine languages. The same conclusion is reached for some distributions for individual works in English, Russian and German presented by Elderton (1949), Fucks (1952) and Herdan (1956). These findings, besides some observations on Fucks' methods will be found in Section 6.4.

6.1.6. Section 6.5 describes the method of fitting lognormal distributions adopted in the present case. This fitting can be done

---

1/ It is, of course, well-known that the lognormal curve fits sentence-length distributions for English prose (Williams, 1940); the same seems to be true for Bengali prose (present study, Chapter 8).

in two ways, first [referred to as LN(a)] by supposing that the observed values of x, viz., 1, 2, ..., etc., represent the intervals 0-1, 1-2, etc., of the underlying continuous variate, and second, [referred to as LN(b)] by supposing that the observed values 1, 2, ..., represent the intervals 0-1.5, 1.5-2.5, etc. Section 6.6 discusses the fit provided by lognormal distributions to Bengali data. The fit was much better than for Fucks' law except for the older works in wholly chaste style. LN(a) was better than LN(b) for works in wholly colloquial style, but LN(b) was superior in the other cases. For both the deviations were often significant but small in the absolute sense. The Kolmogorov distances were often of the order of 2 % or 3 %.

6.1.7. The examination of Poisson and lognormal fits was carried out separately for all the 28 works in Bengali prose subjected to sampling[1] [vide Table 3.1; the short essays and stories were excluded]. Only the "combined" distribution was considered, and not the subsample distributions; also, where both probability and systematic samples were available, the pooled (probability plus systematic) distribution alone was examined. It has been assumed throughout that the word-length distributions are based on random samples. In view of the findings reported in Chapters 2 and 5, this can hardly vitiate any of the conclusions reached.

1/ Fucks (1955) examined one 'average' distribution for each language, a procedure open to serious criticism (vide Section 6.4 below).

6.2.1. Goodness of Fit Criteria : Table 6.1 presents the observed word-length distributions separately for 28 works in Bengali prose, along with the expected distributions based on three different hypotheses, viz., the Fucks law and the two variants of the lognormal hypothesis. Given observed and fitted distributions like those in Table 6.1, one can, of course, compare the values of $p_{obs.}$ and $p_{exp.}$ for each value of x; one can also combine the information obtained for different works by noting (say) how many of the works covered show a particular sign of the difference ($p_{obs} - p_{exp}$), again, separately for each value of x.

6.2.2. • In addition to such methods of examination, three overall measures were used to compare observed and expected distributions of word-length. These indices are presented in Table 6.2. The indices are:

$$\text{(i)}\quad D = \sum_x \left| p_{obs.} - p_{exp.} \right|$$

$$\text{(ii)}\quad K = \sup_x \left| F_o(x) - F_e(x) \right|$$

and

$$\text{(iii)}\quad \chi^2 = \sum_x (f_{obs.} - f_{exp.})^2 / f_{exp.} = n \sum_x \frac{(p_{obs.} - p_{exp})^2}{p_{exp.}}$$

Here $F_o(x)$ and $F_e(x)$ are the observed and fitted distribution functions of x (word-length) n is the number of words in the sample; f_{obs} and f_{exp} are the observed and expected frequencies; also, χ^2 is calculated after suitable pooling of x-classes so that the minimum expected frequency is at least 5 [vide Cochran, 1952].

Table 1: Observed distribution of word-length in syllables (x) along with expected distributions based on three different models[1], separately for 28 selected works in Bengali prose

word-length (syllables)	observed and expected proportions[2] of words											
	Shakuntala				Sitar Vanavas				Durgeshnandini			
	$P_{obs.}$	P_{exp}			$P_{obs.}$	P_{exp}			P_{obs}	P_{exp}		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	1164	1820	883	976	1507	1836	1112	1214	1403	2043	1090	1198
2	3463	3100	4373	4057	3133	3112	4274	3966	3709	3245	4628	4300
3	3376	2642	2881	3033	3427	2637	2684	2823	3243	2576	2713	2869
4	1365	1500	1168	1273	1147	1490	1143	1245	1123	1364	1014	1107
5	445	639	430	442	547	631	457	474	369	541	353	360
6	173	218	160	142	160	214	185	174	89	172	124	113
7	14	62	61	47	53	60	79	64	51	46	47	36
8		19	44	25	13	15	34	24	13	10	18	11
9						3	16	9		2	13	6
10					13	0.6	8	4				
11						0.1	8	3				
no. of sample words		696				750				2359		

1/ The Poisson model states that x-1 obeys the Poisson law. The other two models assume lognormality : LN(a) assumes that the x-values 1,2,.... represent intervals 0-1, 1-2 ..... of an underlying lognormal variate, while for LN(b), the corresponding intervals are 0-1.5, 1.5-2.5, ...

2/ All proportions are on per 10,000 basis.

word-length (syllables)	observed and expected proportions of words											
	Kapalkundala				Visavriksha				Krishnakanter Will			
	P_{obs}	P_{exp}			P_{obs}	P_{exp}			P_{obs}	P_{exp}		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)
1	1582	1930	1397	1506	1587	2325	1395	1520	1964	2594	1641	1780
2	3611	3175	4111	3813	4255	3392	4766	4432	4165	3500	4949	4607
3	2799	2612	2473	2599	2757	2474	2484	2637	2747	2362	2299	2453
4	1095	1432	1113	1212	950	1203	883	964	708	1062	749	818
5	669	589	485	508	337	439	301	306	293	358	238	241
6	203	194	216	209	73	128	105	96	32	97	78	71
7	20	53	102	86	32	31	39	30	24	22	28	20
8	20	12	49	37	4	6	16	10	4	4	11	7
9		3	54	30	4	1	6	2	4	0.7	4	2
10						0.2	5	3		0.1	3	1
11												
no. of sample words		493				2463				2526		

word-length (syllables)	observed and expected proportions of words											
	Anandamath				Devi Choudhurani				Rajsinha			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)	(36)	(37)
1	1649	2354	1345	1471	1983	2838	1533	1683	1585	2205	1343	1462
2	4136	3405	4922	4581	4420	3574	5498	5132	3984	3334	4637	4310
3	2901	2463	2499	2660	2840	2251	2224	2399	2860	2520	2529	2678
4	958	1187	835	911	578	945	561	615	1067	1270	946	1032
5	246	429	264	268	154	298	136	134	358	480	338	347
6	63	124	87	77	15	75	34	29	104	145	124	113
7	31	30	30	22	5	16	10	6	31	37	48	38
8	16	6	11	7			3	1.5	5	8	19	13
9					5	3	0.5	0.3	5	1	9	4
10		1	7	3			0.5	0.2		0.3	7	3
11												
no. of sample words		1910				2007				1930		

word-length (syllables)	observed and expected proportions of words											
	Bouthakuranir Hat				Rajarsi				Chokher Bali			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(38)	(39)	(40)	(41)	(42)	(43)	(44)	(45)	(46)	(47)	(48)	(49)
1	1687	2483	1287	1415	1448	2377	1141	1263	1525	2550	1158	1287
2	4080	3459	5143	4792	4287	3415	5155	4802	4469	3485	5478	5113
3	3233	2409	2500	2671	3124	2453	2608	2785	3042	2381	2509	2697
4	661	1119	759	830	849	1175	786	859	789	1084	649	711
5	294	390	217	217	211	422	219	219	137	370	154	152
6	45	109	64	56	60	121	62	54	30	101	38	32
7		31	30	19	17	29	20	13	8	23	10	6
8						7	6	4		5	4	2
9							2	0.6				
10					4		0.5	0.3				
11							0.5	0.1				
no. of sample words	2419				2321				1318			

Table 6.1 : (Contd.)

word-length (syllables)	observed and expected proportions of words											
	Gora				Chaturanga				Ghare Baire			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(50)	(51)	(52)	(53)	(54)	(55)	(56)	(57)	(58)	(59)	(60)	(61)
1	1574	2617	1388	1523	1683	2684	1342	1481	2067	3353	2239	2416
2	4748	3508	5232	4877	4567	3530	5447	5082	5723	3664	5373	5012
3	2683	2352	2407	2579	2915	2322	2378	2559	1594	2002	1805	1956
4	730	1051	700	765	627	1018	624	683	479	729	436	479
5	236	352	192	193	177	335	154	153	110	199	107	106
6	22	94	56	47	17	88	39	32	21	44	28	24
7	4	21	16	12	13	19	11	8	5	8	8	5
8	4	4	6	3								
9		0.8	3	1		4	5	2		1	4	2
10												
11												
no. of sample words		2713				2312				1901		

word-length (syllables)	observed and expected proportions of words											
	Sheser Kavita				Yogayog				Char-Yari Katha			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(62)	(63)	(64)	(65)	(66)	(67)	(68)	(69)	(70)	(71)	(72)	(73)
1	1852	3019	1910	2070	1963	3108	2197	2360	2271	3466	2815	2697
2	5409	3616	5278	4921	5552	3632	5051	4705	5688	3673	5211	4861
3	1951	2165	2047	2204	1601	2122	1942	2087	1422	1946	1701	1838
4	584	864	554	608	666	827	564	617	482	687	422	463
5	144	259	149	148	168	241	166	167	80	182	108	107
6	45	62	42	37	42	56	52	45	46	39	29	25
7	5	12	13	9	8	11	17	13	11	7	9	7
8	5	2	4	2								
9			2	0.7		2	11	6		1	5	2
10		0.4	0.6	0.2								
11	5		0.2	0.06								
above 11			0.2	0.04								
no. of sample words		2019				1187				872		

Table 6.1 (Contd.)

word-length (syllables)	observed and expected proportions of words											
	Birbaler Halkhata				Pallisamaj				Pather Dabi			
	P_{obs}	$P_{exp.}$			P_{obs}	$P_{exp.}$			P_{obs}	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(74)	(75)	(76)	(77)	(78)	(79)	(80)	(81)	(82)	(83)	(84)	(85)
1	2113	2695	1950	2095	2023	2975	1804	1961	2037	2928	1519	1674
2	4381	3534	4756	4423	4820	3607	5322	4963	4479	3596	5706	5333
3	2392	2317	2148	2289	2506	2186	2101	2262	2810	2209	2153	2334
4	692	1013	739	807	393	884	564	618	552	904	486	533
5	307	332	254	259	214	268	149	149	98	278	104	102
6	67	87	91	84	22	65	41	35	12	68	24	19
7	38	19	36	28	11	13	13	9	12	14	6	4
8		4	14	9	11	2	4	2				
9	10	0.6	6	4						3	2	1
10		0.1	6	2		0.4	2	1				
11												
no. of sample words		1041				890				815		

Table 6.1: (Contd.)

word-length (syllables)	observed and expected proportions of words											
	Pather Panchali				Aparajita				Devayan			
	P_{obs}	$P_{exp.}$			P_{obs}	$P_{exp.}$			P_{obs}	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(86)	(87)	(88)	(89)	(90)	(91)	(92)	(93)	(94)	(95)	(96)	(97)
1	1752	2812	1567	1715	1832	2799	1615	1763	1962	3206	2065	2235
2	4906	3567	5355	4994	4752	3564	5276	4918	5595	3647	5378	5017
3	2484	2263	2248	2417	2592	2269	2239	2405	1710	2074	1913	2069
4	647	957	609	667	565	963	628	688	595	787	479	525
5	188	304	159	158	216	307	171	170	120	224	119	119
6	20	77	43	37	32	78	48	42	16	51	32	26
7	4	16	13	9	5	17	15	10	3	10	10	7
8		3	6	3		3	5	3		2	4	2
9						0.5	2	0.7				
10					5	0.07	0.6	0.2				
11						0.004	0.4	0.1				
no. of sample words	2552				1894				3176			

Table 6.1: (Contd.)

word-length (syllables)	observed and expected proportions of words							
	Dristipat				Janantik			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(98)	(99)	(100)	(101)	(102)	(103)	(104)	(105)
1	1667	2478	1569	1702	1594	2745	1666	1810
2	4570	3457	4819	4483	5275	3549	5098	4749
3	2408	2412	2365	2515	2087	2294	2252	2411
4	969	1122	817	893	768	989	685	749
5	296	391	275	280	203	319	203	204
6	68	109	95	86	73	83	63	56
7	13	25	36	27				
8	8	5	14	9				
9						22	33	21
10		1	10	5				
11								
no. of sample words		2363				690		

Table 6.1 : (Contd.)

word-length (syllables)	observed and expected proportions of words							
	Chachakahini				Deshe Videshe			
	$P_{obs.}$	$P_{exp.}$			$P_{obs.}$	$P_{exp.}$		
		Poisson	LN(a)	LN(b)		Poisson	LN(a)	LN(b)
(1)	(106)	(107)	(108)	(109)	(110)	(111)	(112)	(113)
1	1671	3045	1705	1866	1846	3098	1946	2110
2	5629	3621	5549	5180	5562	3630	5340	4982
3	2031	2152	2086	2256	1808	2127	2002	2161
4	514	853	502	551	619	831	522	571
5	129	254	118	116	152	243	135	134
6	13	60	29	24		57	37	32
7	13	12	8	5	13	11	10	7
8		} 2	} 3	} 2		} 2	} 6	} 3
9								
10								
11								
no. of sample words		778				791		

Table 6.1 Goodness of fit of Poisson and lognormal distributions[1] to observed distributions of word-length in syllables, separately for 28 works in Bengali prose.

work	no. of sample words (n)	index D = $\sum \lvert p_o - p_e \rvert$			K : Kolmogorov distance (%) [2]		χ^2 statistic for LN(b)		
		Poisson	LN(a)	LN(b)	LN(a)	LN(b)	d.f.	χ^2	P-value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. Shakuntala	696	0.219	0.200	0.130	6.29**	4.06	4	15.30	0.001-0.01
2. Sitar Vanavas	750	0.163	0.247	0.196	7.46***	5.40*	4	30.07	< 0.001
3. Durgeshnandini	2359	0.228	0.194	0.124	6.06***	3.86**	4	40.79	"
4. Kapalkundala	493	0.144	0.139	0.088	3.15	1.26	4	8.60	0.05 -0.10
5. Visavriksha	2463	0.230	0.114	0.044	3.19*	1.10	4	6.11	0.10 -0.20
6. Krishnakanter Will	2526	0.211	0.169	0.107	4.62***	2.59	4	33.92	< 0.001
7. Anandamath	1910	0.229	0.181	0.111	5.35***	3.20*	4	24.48	< 0.001
8. Devi Choudhurani	2007	0.288	0.221	0.153	6.28***	4.12**	3	48.67	"
9. Rajsinha	1930	0.199	0.143	0.070	4.10	2.02	4	10.15	0.02 -0.05
10. Bouthakuranir Hat	2419	0.289	0.242	0.182	6.63***	4.40***	4	86.89	< 0.001
11. Rajarsi	2321	0.309	0.178	0.108	5.61***	3.30*	4	29.62	"
12. Chokher Bali	1318	0.329	0.208	0.133	6.42***	4.06*	3	23.64	"
13. Gora	2713	0.314	0.107	0.040	2.99*	0.79	4	10.55	0.02 -0.05
14. Chaturanga	2312	0.326	0.181	0.118	5.39***	3.13*	3	32.16	< 0.001
15. Ghare Baire	1901	0.412	0.079	0.143	1.79	3.63*	3	41.31	"

(contd.)

Table 6.2 (Contd.)

work	no.of sample words (n)	index D = $\sum \lvert P_o - P_e \rvert$			K : Kolmogorov distance (%)		χ^2 statistic for LN(b)		
		Poisson	LN(a)	LN(b)	LN(a)	LN(b)	d.f.	χ^2	P-value
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
16. Sheser Kavita	2019	0.360	0.034	0.101	0.73	2.70	3	20.47	<0.001
17. Yogayog	1187	0.384	0.121	0.179	2.67	4.50*	3	40.48	"
18. Char-Yari Katha	872	0.405	0.111	0.174	2.44	4.26	2	26.43	"
19. Birbaler Halkhata	1041	0.190	0.093	0.037	2.12	0.79	3	3.29	0.30 -0.50
20. Pallisamaj	890	0.308	0.139	0.076	2.83	1.63	3	12.72	0.001-0.01
21. Pather Dabi	815	0.297	0.249	0.173	7.09***	4.91*	2	25.54	<0.001
22. Pather Panchali	2552	0.312	0.098	0.027	2.64	0.51	3	6.48	0.02 -0.05
23. Aparajita	1894	0.303	0.124	0.062	3.07	0.97	3	11.12	0.01 -0.02
24. Devayan	3176	0.390	0.067	0.130	1.14	3.05**	3	57.42	<0.001
25. Dristipat	2363	0.223	0.063	0.036	1.50	0.54	4	6.14	0.10 -0.20
26. Janantik	690	0.345	0.054	0.112	1.05	3.10	3	8.85	0.02 -0.05
27. Chachakahini	778	0.402	0.022	0.094	0.46	2.54	2	6.58	0.02 -0.05
28. Deshe Videshe	791	0.387	0.067	0.130	1.20	3.16	2	12.88	0.001-0.01

1/ The Poisson model assumes that x-1 obeys the Poisson law. The other two models assume lognormality : LN(a) supposes that the x-values 1, 2, ... represent the intervals 0-1, 1-2, ..., of an underlying lognormal variate, while for LN(b), the corresponding intervals are 0-1.5, 1.5-2.5, ...,

2/ Single asterisk denotes significance at 5% level, double asterisk at 1% level and triple asterisk at 0.1% level.

6.2.3. The first measure D was used in all cases for rough examination. It is an excellent indicator for descriptive purposes[1/]. It is rather intractable from the sampling theory point of view, and cannot give a simple test of significance. But, in a rough way, we could judge the significance of D from the value of n, the sample size. This is because D is highly correlated with χ^2/n. Fig. 6.1 brings out this correlation. It is based on comparisons between observed distributions for 28 works in Bengali prose and the corresponding lognormal fits obtained by the second approach /LN(b)/. Thus, the same value of D may be taken as significant or otherwise depending on whether n is large or small. But D is more than a test criterion; it is a direct measure of divergence between observed and expected distributions.

6.2.4. For the distributions for Bengali prose the Poisson model was found to be poor even by visual examination and by using the index D. The same is also indicated by the difference between variance of x, viz., s_x^2, and $(\bar{x} - 1)$, where $\bar{x}$ is the mean of x; this difference should vanish under Fucks' hypothesis. /This point was overlooked by Fucks (1955)./

6.2.5. Since the lognormal distribution gave much better fit to the data on Bengali, the measure K, known as the Kolmogorov distance[2/], was employed to examine the fit more closely. This also was used more

1/ Vide Lahiri and Ganguly (1951) for an application of this measure, where its practical advantages are brought out fully.

2/ For a general account of the Kolmogorov statistic, vide Kendall and Stuart, Vol. 2, 1961, Chapter 30 and also Siegel, 1956.

as a measure of divergence than as a test criterion, because in the present case, it can only lead to very rough tests of significance. Apart from the deviations from random sampling, which are not at all consequential, this statistic suffers from two serious limitations in the present situation.

6.2.6. First, the Kolmogorov statistic strictly applies to continuous variates; when applied to discrete data it gives conservative tests which err in the "safe" direction, that is, tests with the true level of significance less than the stated value (Scheffé, 1943; Noether, 1963). Second, the sampling distribution of the Kolmogorov statistic is not known for the case where some parameters have been estimated from the sample : There is some evidence indicating that the resulting probability statements probably err on the "safe" side (Massey, 1951), but there is no exact result analogous to that for χ^2. In the present case, the distribution is confined to a few discrete points and two parameters are estimated from the observed word-length distributions. This explains why the K-statistic is actually found to be less sensitive than the χ^2-test, contrary to what is generally said about the relative efficiencies of the two tests for continuous distributions where the χ^2-test involves grouping of observations and loss of information.

6.2.7. Out of the two methods of fitting the lognormal distribution the second, denoted by LN(b), gave a somewhat better fit, on

the whole. Chi-Square was applied as a final test of significance[1] for judging whether the observed distributions differ significantly from lognormal distributions fitted by this second method. The small deviations from random sampling were ignored, and also that the method of estimating the parameters of the lognormal distribution [vide Section 6.5] was not fully efficient, as required for justifying the usual rule for getting the degrees of freedom for χ^2.

6.3.1 Inadequacy of Poisson law for Bengali Prose : Table 6.1 presents, among other things, the observed word-length distributions for different works and the expected distributions assuming the Fucks law. If x-1 is a Poisson variate with parameter λ, the maximum likelihood estimate of λ is $\bar{x}$-1, where $\bar{x}$ is the sample mean of x. This estimate was used for finding the expected distribution. It can be seen that for most of the works the Poisson fit is not at all satisfactory.

6.3.2 Col. (3) of Table 6.2 shows the index D for the goodness of the Poisson fit. The observed and the fitted distributions are also shown in Fig. 6.2 for six representative works. For each work, the observed word-length distribution is given by a frequency polygon and the fitted Poisson distribution by means of crosses; the lognormal fits LN(a) and LN(b) are also shown in the same figure.

1/ χ^2/n could be used as a measure of goodness of fit.

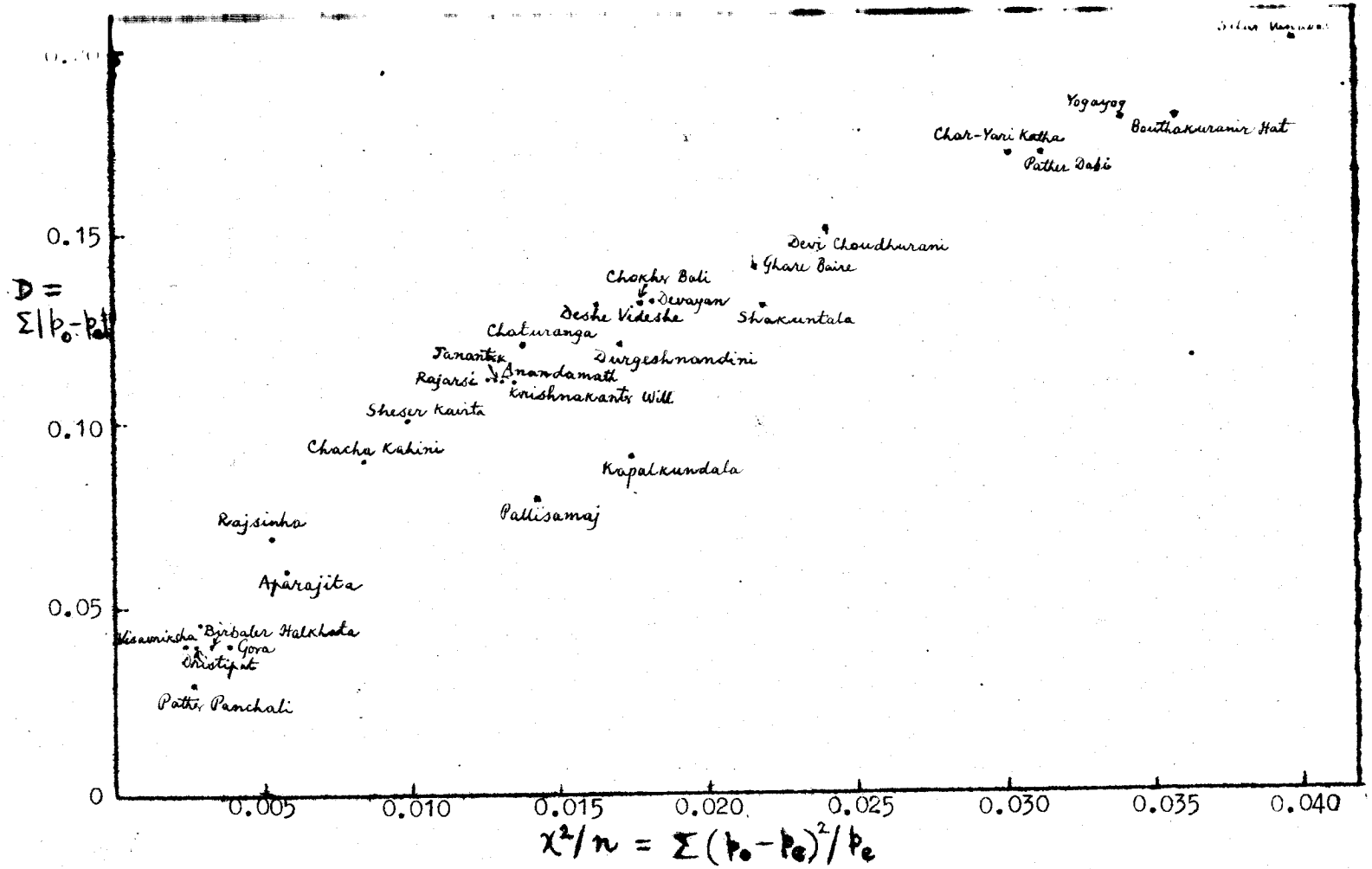


Fig.6.1: Correlation between two measures of goodness of fit of lognormal distributions (fitted by the second approach i.e., LN(b)) to the estimated distributions of word-length in syllables, separately for 28 works in Bengali prose [Vide paragraphs 6.2.1-3 for explanation]

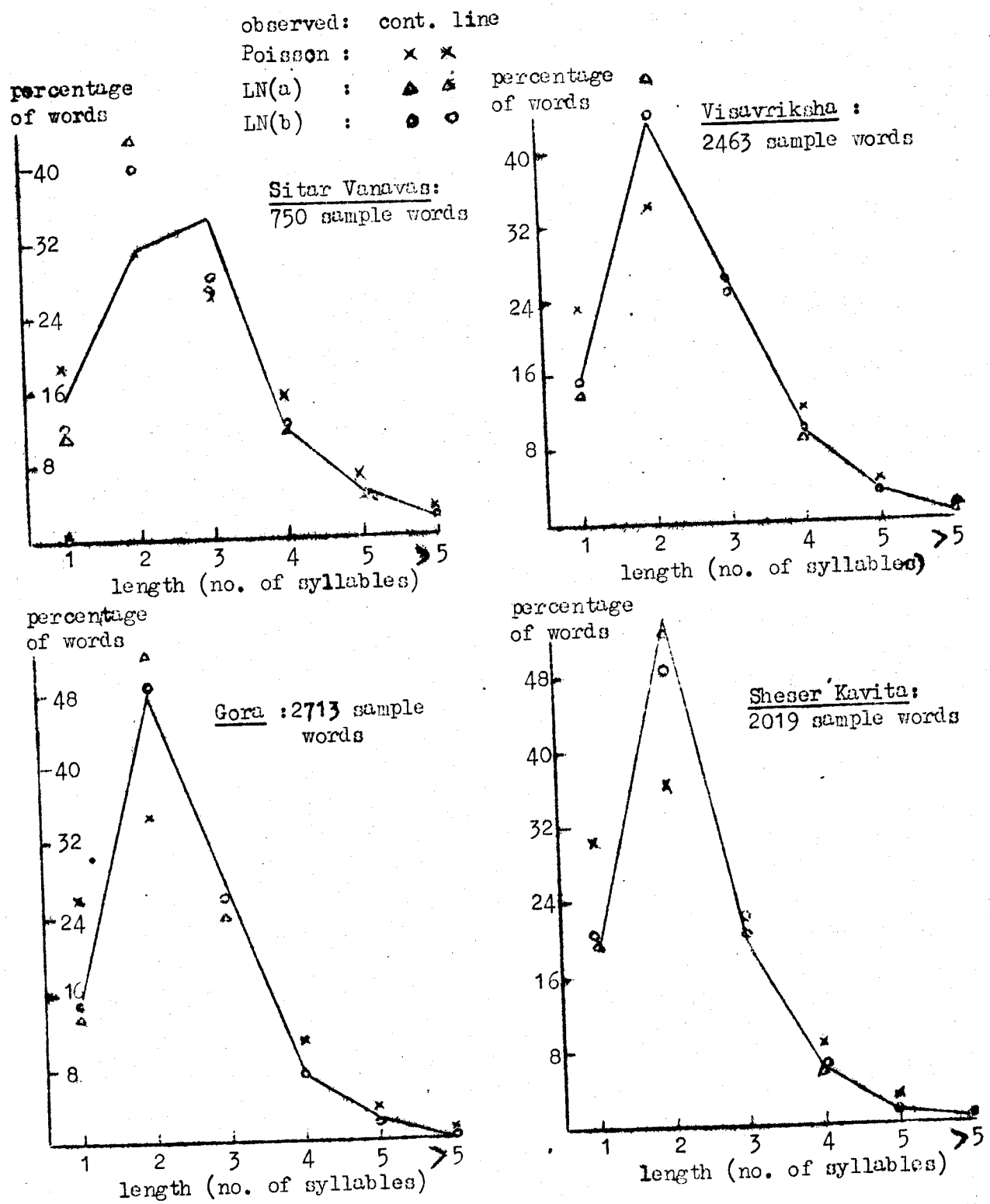
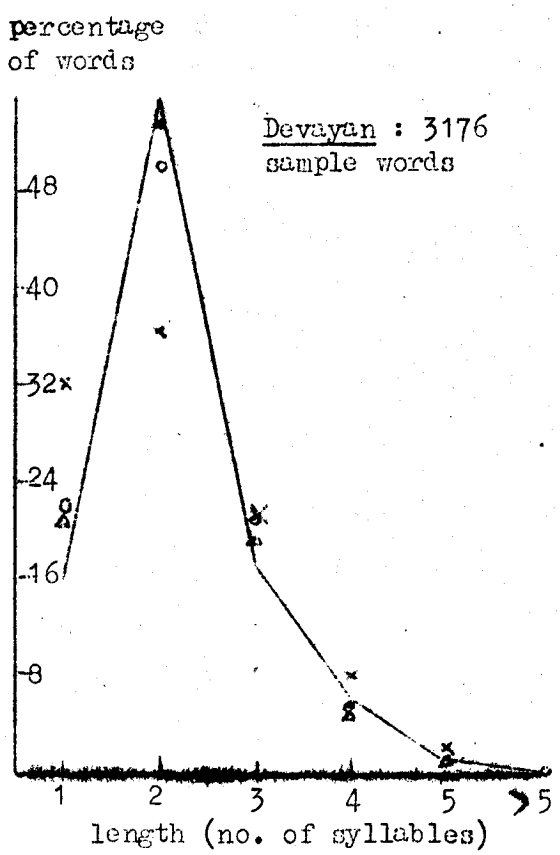


Fig. 6.2: Observed distributions of word-length in syllables (x) along with fitted distributions (i) assuming x-1 is a Poisson variate (ii) assuming x =1,2,3,... regresents intervals 0-1,1-2,2-3,.... or 0-1.5, 1.5-2.5, 2.5-3.5,.... of an underlying lognormal variate [LN(a) or LN(b)] separately for six selected works in Bengali prose [vide Section 6.1].


percentage
of words
Devayan : 3176
sample words
48
40
32
24
16
8
1
2
3
4
5
>5
length (no. of syllables)

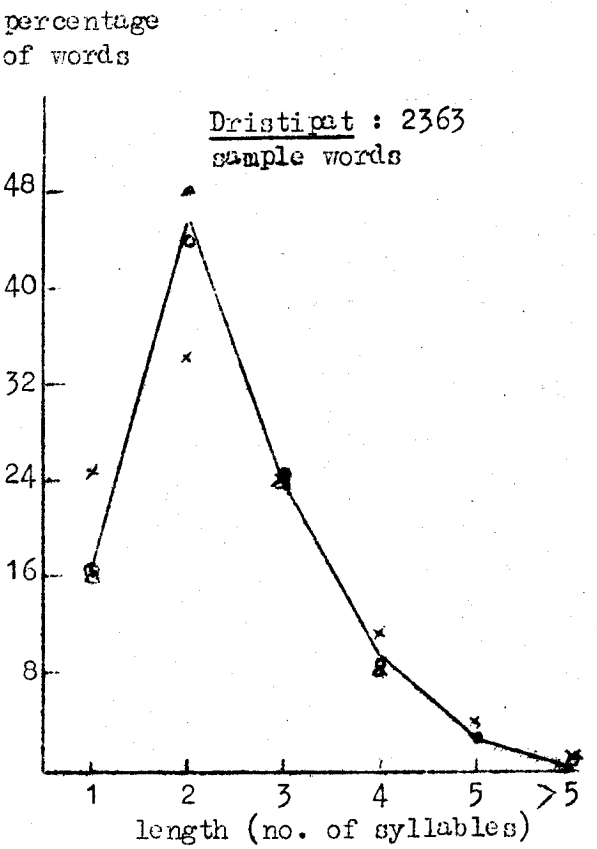

percentage
of words
Dristipat : 2363
sample words
48
40
32
24
16
8
1
2
3
4
5
>5
length (no. of syllables)

Fig. 6.2: (contd.)

6.3.3. The Poisson fit seems to be a coarse first approximation for the older works in chaste style by Vidyasagar and Bamkimchandra (nos. 1-7 and 9 in Table 6.2), the values of D being of the order of 0.15 to 0.20. A few later works (numbered 19 and 25) fall near this range, which may be ascribed to heterogeneity of these works and other factors. Other works have D around 0.30, which means more serious divergence between observations and the Poisson fit. These include many works employing the chaste style outside conversations and the colloquial style inside, viz., those numbered 10-14, 20-23 in Table 6.2. "Devi Choudhurani" (no.8) falls in this group; this is in the chaste style but employs a nearly colloquial style in conversations. Finally, D is nearly 0.35 or 0.40, and the fit is extremely poor, for several works (nos. 15-18, 24, 26-28) most of which are entirely in the colloquial style. All these categories of works are represented in Fig. 6.2 1/.

6.3.4. Generally speaking, p_o falls below p_e for x=1, by upto 15%, approximately, but above p_e for x=2 by even upto 20%; for x=3, the difference is seen to have either sign, for $x > 3$, the values of p_o tend to be lower than p_e .

6.3.5. The inadequacy of the Poisson fit is also evident from Table 3.1 in Chapter 3 : The sample variance s_x^2 is appreciably less than $\bar{x} - 1$, where $\bar{x}$ is the sample mean, for most of the 28 works. The exceptions are "Sitar Vanavas", "Kapalkundala" and "Birbaler Halkhata" where the differences are rather small [vide also "Samya"].

1/ The fact that D depends on sample size may be ignored for such rough grouping.

6.3.6. It is needless to carry out formal tests of significance, as the Poisson fit is evidently quite poor. It will be seen that the lognormal fit is generally closer. Goodness of fit was examined in greater detail for the lognormal hypotheses, and even there the χ^2-test gave highly significant results.

6.4.1. Poisson law for other languages : Some time was spent in examining, in a similar manner, how far the Poisson fit is really adequate for the eight languages as claimed by Fucks (1955). We shall breifly report on our findings here, without presenting any tables and graphs.

6.4.2. The situation is better than for Bengali prose, for the value of D is 0.03 for Esperanto, 0.08 for German, and of the order of 0.10 or 0.15 for the six remaining languages. (For Arabic, D is 0.31.) Fucks does not present any figure for sample size, so that tests of significance are impossible. Presumably, however the sizes are quite large. If it is so, the fit cannot be said to be perfectly satisfactory, even for the eight languages, excepting Arabic.

6.4.3. The difference $s_x^2 - (\bar{x} - 1)$ is well below zero for Arabic, Latin and Turkish; in the neighbourhood of zero for Esperanto and German; and fairly above zero for the four remaining languages. Some values of $\bar{x}$ given by Fucks seem to be misprints. We may quote some figures which should be correct : Arabic - $s_x = 0.811$, $\bar{x} = 2.104$; Turkish - $s_x = 1.081$, $\bar{x} = 2.459$; Esperanto - $s_x = 0.925$,

$\bar{x}$ = 1.895; German -s_x = 0.863, $\bar{x}$ = 1.634; Russian -s_x = 1.191, $\bar{x}$ = 2.230; English -s_x = 0.751, $\bar{x}$ = 1.406.

6.4.4. There is, moreover, an important objection to Fucks' methodology. Fucks used one "average" distribution $\{ P_x ; x=1,2,3, \ldots\ldots \}$ for each language. This 'average' distribution was obtained by taking the unweighted average of corresponding values of P_x for different works in the given language. Fucks recommended the use of a sufficiently large number of works for such averaging, so that the addition of one extra work does not affect the averages appreciably.

6.4.5. This concept of an "average" distribution for a language is ill-defined and impractical. Each language has different fields of literature, each employing a characteristic level of (average) word-length. Unless the relative weightages of different types of works are specified, the average distribution will depend greatly on the relative weightages actually given in the sampling of works. Thus, Fucks' data on English gives $\bar{x}$ =1.406, as against 1.43 given by Dewey (1923) on the basis of a representative sample of modern English prose.

6.4.6. It is therefore desirable to study the distributions for individual works instead of any pooled or average distributions for languages. The Fucks model cannot be applicable in both cases. Thus, if individual works obey the Poisson law, the average distribution will have variance larger than the mean, because of the between works variation in the average of word-length.

6.4.7. We therefore tried the Poisson model for some individual works in English, German and Russian for which the word-length distributions are presented in Fucks (1952), Elderton (1949) and Herdan (1956). In most cases, s_x^2 exceeds $(\bar{x}-1)$ by an appreciable margin.

6.4.8. As regards English works, Gray's Poems (D = 0.02) and Genesis (D = 0.03) show excellent agreement; works by Shakespeare, Galsworthy, Huxley and Gray's Letters gave D of the order of 0.10 or 0.15; lastly, for works by Macaulay, Carlyle, Johnson, Bacon and Gibbon — all showing higher $\bar{x}$-values — the values of D were above 0.20, usually in the neighbourhood of 0.25. Generally speaking, the observed frequency is smaller at x = 2 but larger for other values of x.

6.4.9. As regards German works, Rilke's Cornet gives D = 0.03, and Goethe's Wilhelm Meister D = 0.04; works by Carossa, Hesse, Mann and Lichtenberg show D of the order of 0.10 to 0.15; but Jaspers with D = 0.30 showed very serious divergence from the Poisson law. On the whole, the German works show too many polysyllabled words ($x > 4$) for the Fucks model to fit the data.

6.4.10. For three of the four Russian works covered in Herdan (1956), the index D is a little above 0.10; the remaining one, Turgeneff's Rudin gave D = 0.17.

6.4.11. We may quote the values of s_x^2 and $\bar{x}$ for some of the works mentioned in the foregoing paras : Gray's Poems — $s_x^2 = 0.338$, $\bar{x} = 1.326$; Genesis — $s_x^2 = 0.238$, $\bar{x} = 1.215$; Shakespeare's Othello — $s_x^2 = 0.372$, $\bar{x} = 1.285$; Macaulay's Essays — $s_x^2 = 0.86$, $\bar{x} = 1.55$; Rilke's Cornet — $s_x^2 = 0.449$, $\bar{x} = 1.463$; Goethe's Wilhelm Meister — $s_x^2 = 0.764$, $\bar{x} = 1.734$; Mann's Buddenbrooks — $s_x^2 = 0.900$, $\bar{x} = 1.777$; Jaspers' Der Phil — $s_x^2 = 1.354$, $\bar{x} = 1.886$; Tolstoy's Autobiography — $s_x^2 = 1.415$, $\bar{x} = 2.232$; Turgeneff's Rudin — $s_x^2 = 1.260$, $\bar{x} = 2.230$.

6.4.12. As already stated, Fucks (1952) does not present sample sizes, but presumably the sizes were large; the samples taken by Elderton (1949) were also large, generally speaking; only those employed by Herdan (1956) were rather small. It is apparent that many of the above-mentioned works would show significant and appreciable deviations from Fucks' model, while a few would show satisfactory agreement.

6.4.13. No attempt was made to fit the more general models put forward by Fucks (1955, pp.163- 7). As admitted by Fucks himself, they are artificial and can fit any word-length distribution if one includes a sufficiently large number of parameters. The stochastic schemes for generating these distributions or even the Poisson distribution are not as realistic or suggestive as stochastic schemes often are and ought to be. /Vide in this connection comments by Quastler during discussion on Fucks' paper./

6.5.1. Fitting Lognormal Distributions : When the Poisson distribution generally gave a poor fit, we tried the lognormal distribution on the word-length distributions for Bengali prose.

6.5.2. The lognormal distribution is continuous, with the variable ranging from 0 to ∞, while the observed distributions are discrete, with x assuming the values 1, 2, 3; etc. Fitting lognormal curve to such discrete variates should be done on the assumption that the observed discrete variate is the manifestation of an underlying continuous variate, which is lognormal, and that the observed frequencies at 1, 2, 3, ... are really the grouped frequencies of the continuous distribution for the intervals 0-1, 1-2, 2-3, ... respectively [Aitchison and Brown, 1957, pp.92-3, 103-4; Williams, 1956; Herdan, 1958[1/]]. This approach was followed in the present study also, and this variant of the lognormal hypothesis has been denoted by LN(a) in Tables 6.1 and 6.2.

6.5.3. Let x denote the observed word-length variate, and let the underlying continuous variate x' be lognormal with $E(\log_e x') = \theta$ and $Var(\log_e x') = \lambda^2$. The first problem was to estimate the parameters θ and λ for the different word-length (x) distributions, considered as grouped versions of the distribution of 'x'.

6.5.4. Maximum likelihood estimation was obviously too laborious, as it involved an iterative procedure, since the intervals are not

1/ Herdan examined the length-distribution of relatively frequent words, which would be appreciably different from the distribution of all words.

equal on the logarithmic scale; and there were many word-length distributions to be treated separately. The method of quantiles seemed to be the only objective method which is both convenient and reasonably efficient in general (Aitchison and Brown, 1957, Chapter 5).

6.5.5. The method of quantiles consists in choosing two values of x, denoted by x_1 and x_2, (say 2 and 5 in the present case,) and estimating θ and λ from the two equations

$$\log_e x_i = \theta + \lambda t_{P_i} \quad (i=1, 2)$$

where P_1 and P_2 are the cumulative proportions of observed x-values less than or equal to x_1 and x_2, respectively, and t_{P_1} and t_{P_2} the corresponding standard normal deviates defined by

$$P_i = \int_{-\infty}^{t_{P_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (i = 1,2)$$

Aitchison and Brown (1957, pp.40-42) recommend that for maximum efficiency of the estimate $\hat{\theta}$, x_1 and x_2 should be chosen near the 27% and the 73% percentiles of the distribution, while for maximum efficiency of the estimate $\hat{\lambda}$, x_1 and x_2 should be near the 7% and the 93% points.

6.5.6. In the present case, the method was adopted in a modified form. One reason for modification was that since the relative frequencies of x=1, 2, 3 and 4 were usually of the order of 20%, 40%,

20% and 10% respectively, x_1 and x_2 could not be chosen symmetrically and/or near the optimum percentiles as mentioned above. Secondly, while the ogive on log-probit scale [vide Figs. 6.3(a) and 6.3(b)] was often approximately linear, appreciable curvature was noticed in several cases. The ordinary method of quantiles would be rather hazardous in these latter-mentioned cases, for it is equivalent to a method of selected points on the log-probit diagram.

6.5.7. The method finally adopted was this : We found the cumulative proportions of observed x-values below and upto x=1,2,3 and 4. These may be denoted by P_1, P_2, P_3 and P_4 respectively (note a change in symbols). The corresponding probits or normal deviates were then found out; denote these by t_{P_1}, t_{P_2}, etc. The following equations were solved for estimating both θ and λ :

$$\log_e 1 + \log_e 2 = 2\theta + \lambda (t_{P_1} + t_{P_2})$$

and $$\log_e 3 + \log_e 4 = 2\theta + \lambda (t_{P_3} + t_{P_4})$$

This amounts to passing a straight line on the log-probit graph through the point lying "midway" between the observed points for x = 1 and x = 2 and the point lying "midway" between the observed points for x=3 and x=4. Most of the word-length distributions are largely confined to these four points. This seems to be a reasonably efficient method of fitting.

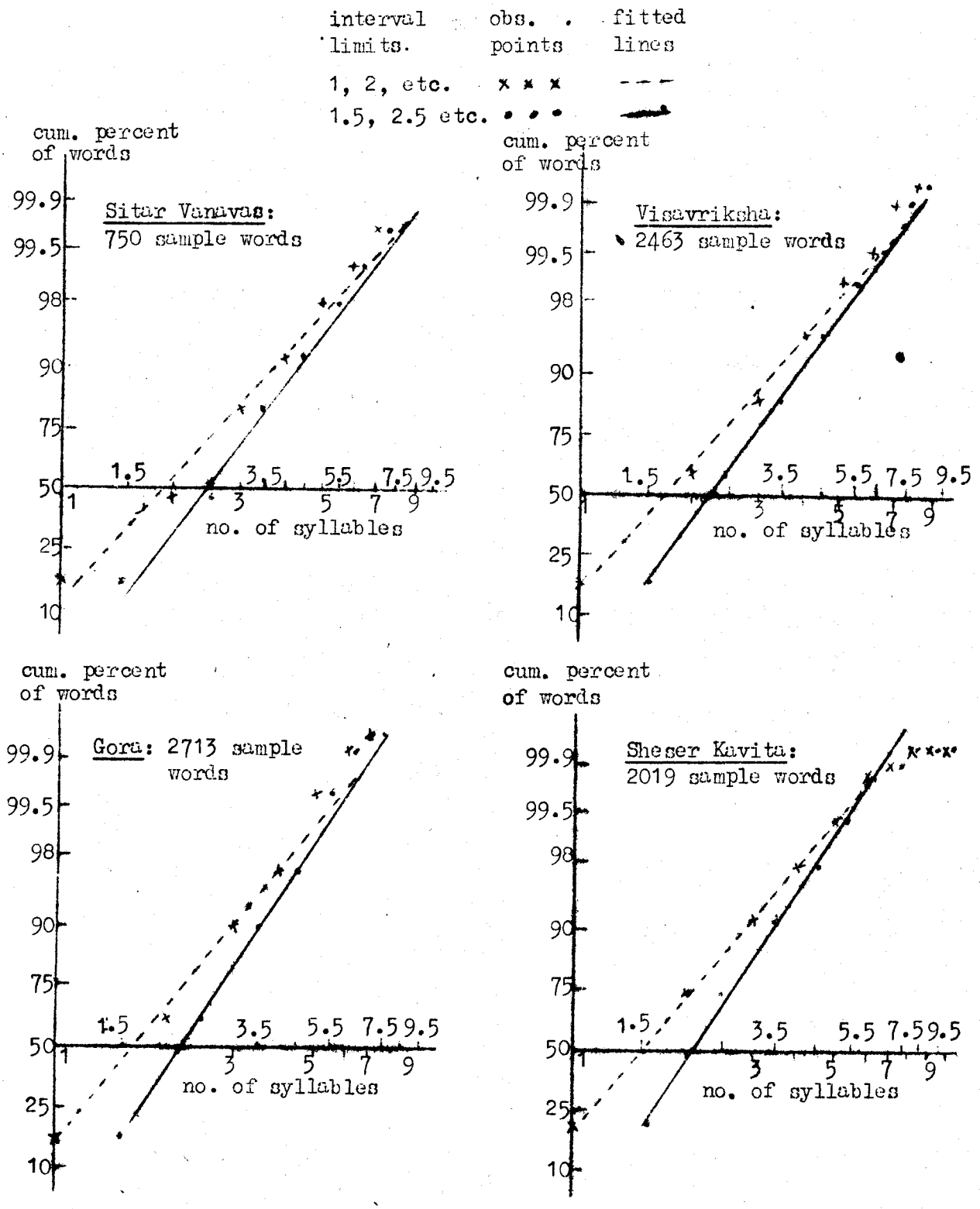
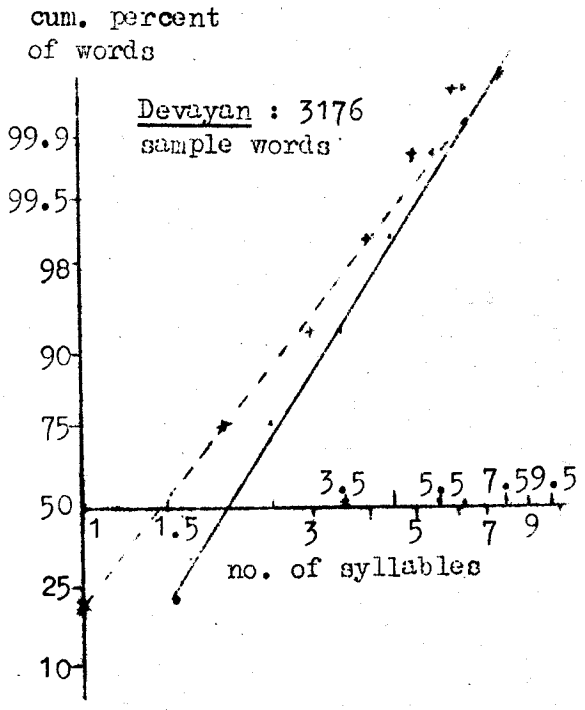


Fig.6.3: Ogives on log-probit scale for distributions of words by length in syllables for six selected works in Bengali prose.


cum. percent
of words
Devayan : 3176
sample words
99.9
99.5
98
90
75
50
25
10
1
1.5
3
3.5
5
5.5
7
7.5
9
9.5
no. of syllables

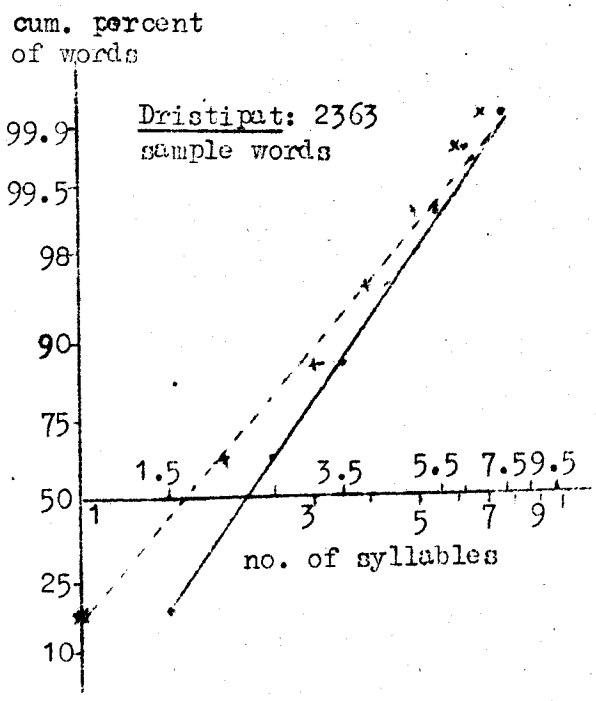
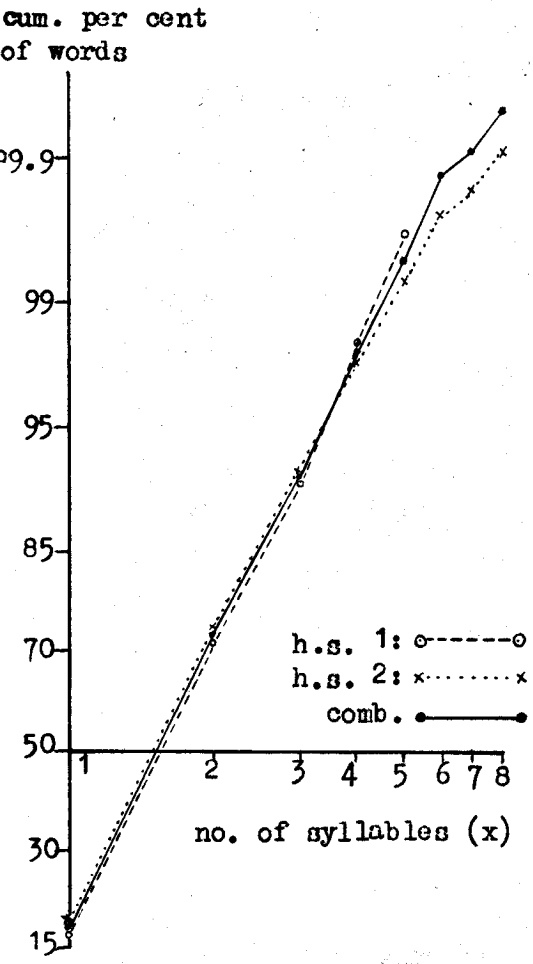

cum. percent
of words
Dristipat: 2363
sample words
99.9
99.5
98
90
75
50
25
10
1
1.5
3
3.5
5
5.5
7
7.5
9
9.5
no. of syllables

Fig. 6.3 : (contd.)



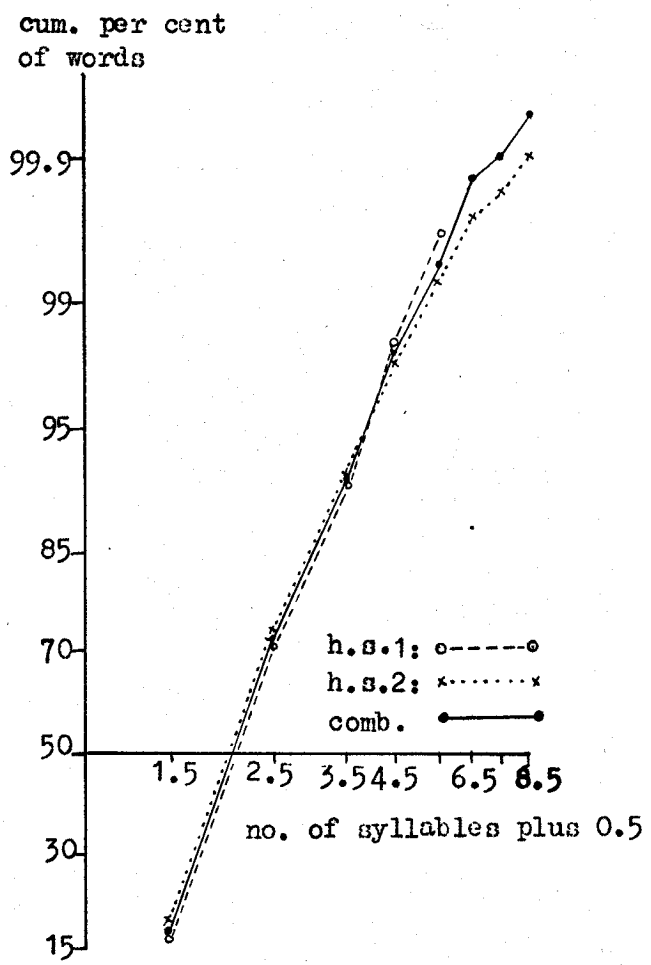


Fig. 6.3(a): Examination of lognormality of the distribution of words by length in syllables (x), estimated from the pooled sample of 2019 words from "Sheser Kavita", assuming that x =1,2,3,... represents (i) intervals 0-1, 1-2, 2-3, ... and (ii) 0-1.5, 1.5 - 2.5, 2.5—3.5 .... of the underlying lognormal variate.

6.5.8. The curvature of the log-probit graphs [Fig. 6.3] suggested a second approach to fitting the lognormal curve: We assume now that the observed frequencies at x = 1,2,3 etc., are really the grouped frequencies of the underlying continuous distribution for the intervals 0-1.5, 1.5 - 2.5, 2.5 - 3.5, etc. One would then solve the equations

$$\log_e 1.5 + \log_e 2.5 = 2\theta + \lambda (t_{P_1} + t_{P_2})$$

$$\text{and} \quad \log_e 3.5 + \log_e 4.5 = 2\theta + \lambda (t_{P_3} + t_{P_4})$$

This variant of the lognormal hypothesis has been denoted in the present chapter by LN(b). Both the approaches LN(a) and LN(b) were followed for fitting lognormal distributions to each of the 28 word-length distributions for Bengali prose.

6.5.9. The estimates of θ and λ are shown in Table 6.3. The observed and expected distributions are shown in Table 6.1, and the measures of goodness of fit in Table 6.2. Fig. 6.2 shows the fit for six representative works by means of frequency polygons. Fig. 6.3 shows the same by means of ogives on log-probit scale.

6.5.10. One comment about the second approach LN(b). It would probably have been more logical to consider the class-limits as $0, \sqrt{1\times2}, \sqrt{2\times3}, \sqrt{3\times4}$, etc., instead of 1.5, 2.5, 3.5, ... This was tried in one case and the results are shown in Table 6.4. The difference is not negligible, but cannot affect the broad conclusions reached in our study.

Table 6.3: Estimates of parameters of lognormal distributions fitted to observed distributions of word-length in syllables, separately for 28 works in Bengali prose*.

work	no. of sample words	estimates of lognormal parameters			
		LN(a)		LN(b)	
		θ	λ	θ	λ
(1)	(2)	(3)	(4)	(5)	(6)
1. Shakuntala	696	0.2874	0.2127	0.3966	0.1703
2. Sitar Vanavas	750	0.2788	0.2286	0.3898	0.1830
3. Durgeshnandini	2359	0.2624	0.2130	0.3766	0.1706
4. Kapalkundala	493	0.2692	0.2489	0.3821	0.1993
5. Visavriksha	2463	0.2365	0.2185	0.3558	0.1750
6. Krishnakanter Will	2526	0.2121	0.2170	0.3363	0.1737
7. Anandamath	1886	0.2329	0.2107	0.3530	0.1687
8. Devi Choudhurani	2007	0.1978	0.1935	0.3249	0.1549
9. Rajsinha	1930	0.2458	0.2222	0.3633	0.1779
10. Bouthakuranir Hat	2419	0.2274	0.2008	0.3486	0.1608
11. Rajarsi	2321	0.2362	0.1960	0.3556	0.1570
12. Chokher Bali	1318	0.2225	0.1859	0.3446	0.1489
13. Gora	2713	0.2174	0.2002	0.3405	0.1603
14. Chaturanga	2312	0.2120	0.1916	0.3362	0.1534
15. Ghare Baire	1901	0.1555	0.2049	0.2910	0.1641
16. Yogayog	1187	0.1698	0.2196	0.3025	0.1759
17. Sheser Kavita	2019	0.1811	0.2071	0.3115	0.1659
18. Char-Yari Katha	872	0.1421	0.2122	0.2803	0.1699
19. Birbaler Halkhata	1041	0.1989	0.2313	0.3257	0.1853
20. Pallisamaj	890	0.1865	0.2041	0.3158	0.1634
21. Pather Dabi	815	0.1912	0.1859	0.3196	0.1489
22. Pather Panchali	2552	0.2009	0.1993	0.3274	0.1596
23. Aparajita	1894	0.2008	0.2032	0.3273	0.1627
24. Devayan	3176	0.1670	0.2040	0.3002	0.1634
25. Dristipat	2363	0.2225	0.2210	0.3447	0.1769
26. Janantik	690	0.2044	0.2112	0.3301	0.1691
27. Chacha Kahini	778	0.1848	0.1941	0.3144	0.1554
28. Deshe Videshe	791	0.1763	0.2047	0.3077	0.1639

* LN(a) supposes that the observed lengths 1,2, ... represent intervals 0-1, 1-2, ... of the underlying lognormal variate; for LN(b), the corresponding intervals are 0 -1.5, 1.5 - 2.5, ... In either case, θ and λ are the mean and the s.d. of logarithms of the underlying variate.

Table 6.4 : Comparison between lognormal fits to word-length distribution for "Pather Panchali" (i) with the class-limits as 1.5, 2.5 etc. and (ii) with the class-limits as $\sqrt{2}$, $\sqrt{6}$, etc., based on the pooled (probability plus systematic) sample from "Pather Panchali": 2552 words.

word-length in syllables	proportion of words		
	observed	'expected' (p_e) under lognormal hypotheses	
	(p_o)	class-limits 1.5, 2.5, ...[a]	class-limits $\sqrt{1}$, $\sqrt{2}$, ...[b]
(1)	(2)	(3)	(4)
1	0.1752	0.1715	0.1659
2	0.4906	0.4994	0.5120
3	0.2484	0.2417	0.2361
4	0.0647	0.0667	0.0649
5	0.0188	0.0158	0.0159
6	0.0020	0.0037	0.0038
7	0.0004	0.0009	0.0010
above 7		0.0003	0.0004
total	1.0000	1.0000	1.0000
$\Sigma\lvert p_o - p_e \rvert$		0.0267	0.0489
K-statistic		0.0051	0.0121

a/ $\theta = 0.3274$, $\lambda = 0.1596$ b/ $\theta = 0.3121$, $\lambda = 0.1665$

[N.B.: $\theta = E(\log_{10} x')$ and $\lambda^2 = \text{Var}(\log_{10} x')$, x' being the underlying continuous variate.]

6.6.1. The Goodness of the Lognormal fit : Table 6.2 sets out all the criteria of goodness of fit employed to examine the goodness of the Poisson and the lognormal fits. The first thing to note is that the index D is much smaller, on an average, for the lognormal fit than for the Poisson fit, and that the two approaches to fitting the lognormal are about equally good or equally bad, on the whole. The average values are 0.293 for the Poisson fit, 0.137 for LN(a) and 0.110 for LN(b). But the picture is very different for different works. As we shall see, this is partly why the present chapter cannot lead to a single conclusion for Bengali prose in general.

6.6.2. First of all, the Poisson fit seems to be the best of the three for 'Sitar Vanavas' (work no.2 of Table 6.2); generally speaking, the Poisson fit is not much poorer than the lognormal fit for the older works in chaste style. With the passage of time, as the colloquial style became more and more dominant, the Poisson fit became poorer and poorer, and the superiority of the two lognormal fits more and more conspicuous.

6.6.3. We next compare the two methods of fitting the lognormal curve. Several interesting points may be noticed :

(1) Generally speaking, LN(a) gives closer fit for works wholly in the colloquial style, e.g., "Ghare Baire" and "Chacha-Kahini", while LN(b) is superior for "Shakuntala" etc., where the chaste style is used at least in conversations. There are a few

exceptions to this general observation, e.g., "Birbaler Halkhata", but these may be explained away. In all, for 20 out of 28 works, the approach LN(b) was found to be superior.

(2) The difference between the two values of D tends to be surprisingly close to 0.06 or 0.07, whichever method of fitting the lognormal curve happens to be superior.

(3) The selection of 28 works gave higher weightage to works employing the older chaste style than to modern works wholly in the colloquial style. One should not therefore be unduly influenced by the overall superiority of LN(b) over LN(a).

(4) LN(a) gives D about 0.15 or 0.20 for older works but only about 0.1 or even 0.05 for the later works in colloquial style; LN(b) hardly shows any time-trend in D, which fluctuates around 0.1, roughly speaking.

6.6.4. So far we compared the three fits by means of the index D, depending partly on intuition for ideas about significance. We now consider the tests of significance, which place us on firmer ground. The Kolmogorov test was applied for both methods of fitting lognormal distribution, and the χ^2-test, for the second method, i.e., LN(b).

6.6.5. Consider first the Kolmogorov distance K. For LN(a), the values are somewhat large, of the order of 5% for the earlier works, but only about 2%, on an average, for the later works, starting from "Gora" by Tagore. Compared with LN(a), LN(b) gives smaller K for the

earlier works and larger values for the later ones — just as in the case of D. On the whole, the values may be said to fluctuate around 3%, without showing any clear time-trend.

6.6.6. The foregoing paragraph ignores the fact that K tends to decrease with the size of sample; but it serves to stress that, on the whole, the lognormal fits are fairly close, even though the deviations are frequently significant in the statistical sense.

6.6.7. The Kolmogorov test shows significance at the 5% level in 14 cases for LN(a), and in 12 cases for LN(b); at the 1% level in 12 and 4 cases respectively; and at the 0.1% level in 10 and 1 cases respectively. Any combination of the 28 tests would obviously give a highly significant result even for the LN(b) hypothesis. And this inspite of the fact stressed in Section 6.2, viz., that the K-test is somewhat conservative and also insensitive in the present circumstances.

6.6.8. The χ^2-test shows significant deviations between observed distributions and the LN(b) hypothesis in most cases. Among the exceptions are "Kapalkundala" and "Birbaler Halkhata" where the sample sizes are not sufficiently large to show up the differences as significant; in the remaining cases, e.g., "Dristipat", the fit seems to be rather close, D being nearly 0.04 or even smaller. The total of the 28 χ^2's is 680.68, which, for a χ^2 with 92 d.f.'s, is a remarkably high value ($\sqrt{2\chi^2} - \sqrt{2n-1}$ is over 23), far beyond

any ordinary level of significance. Also, as many as 15 of the 28 χ^2's are significant at the 0.1% level. There is thus, no doubt, whatsover, about the significance of the deviations from the LN(b) model.

6.6.9. One should not, however, be too seriously discouraged by the generally significant results of the χ^2-test. It still appears that the LN(b) is a first approximation to the true model. This is indicated by the Kolmogorov distances discussed in earlier paragraphs which seem to be reasonably small in an absolute sense. Unless one hits upon a perfect probability model — which one can very seldom do in practice, at least in cases like the present one— there will surely be a significant goodness of fit χ^2, provided only that the sample size is sufficiently large. For χ^2 can be written as

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = n \sum \frac{(p_o - p_e)^2}{p_e}$$

and the factor $\sum \frac{(p_o - p_e)^2}{p_e}$ is like $\sum |p_o - p_e|$, a measure of divergence between observations and theory [vide Fig. 6.1 in this connection]. For the same divergence between observed and fitted distributions, χ^2 would increase proportionately to sample size. The many significant results in the present case are really due to the relatively large sample sizes.

6.6.10. The average value of $\sum \frac{(p_o - p_e)^2}{p_e}$ is found to be 0.01612; so if n were 500, this much divergence would give a χ^2 nearly about 8.06, which for 3 or 4 d.f.'s nearly significant at the 5% level. So the approximation furnished by LN(b) is such that deviations come out significant by the χ^2-test when sample size is 500 or more. The same may obviously be said about LN(a) also.

6.6.11. One final onservation before concluding this section. The K-test showed that for both LN(a) and LN(b), the maximum distance between $F_o(x)$ and $F_e(x)$ occurred frequently at x = 2. The fitted value was the larger one, generally speaking, for works employing the chaste style [LN(b) better]; but the fitted value was usually smaller for works entirely in the colloquial style [LN(a) superior].

6.6.12. This inevitably reminds us of what is called by linguists the law of bimorism for Bengali [Chatterjee, 1945 , p. 38] : All words in the colloquial form tend to have two syllables or multiples thereof. This results in a much higher proportion of bisyllabled words in colloquial Bengali than found in chaste Bengali. The fit given by theoretical distributions seems to depend, very critically, upon how well the model can predict the proportion at x=2.

6.7.1. Concluding Observations : The lognormal fit might be carried out separately for conversational and other words. In view of the findings of Chapters 4 and 5, this should improve the fit by increasing the homogeneity of the observed data. This expectation

was not fulfilled, however, in the trial with the pooled-sample data for "Gora". Table 6.5 shows the results. For all words considered together, the index D was 0.107 for LN(a) and 0.040 for LN(b) [vide Table 6.2]; the K-distances were 2.99 and 0.79 respectively.

Table 6.5: Lognormal fits for observed distributions of word-length in syllables, separately for conversational and other words, based on the pooled (probability plus systematic) sample of 2713 words from Gora[1].

word-length (syllables)	obs.		exp.			
			LN(a)		LN(b)	
	conv. words	other words	conv. words[2]	other words[3]	conv. words[4]	other words[5]
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.2058	0.1275	0.2065	0.1023	0.2242	0.1136
2	0.5527	0.4267	0.5512	0.5027	0.5142	0.4678
3	0.1787	0.3236	0.1860	0.2730	0.2020	0.2907
4	0.0512	0.0864	0.0430	0.0862	0.0472	0.0944
5	0.0116	0.0310	0.0099	0.0250	0.0098	0.0250
6		0.0036		0.0073		0.0063
7		0.0006	0.0034	0.0023	0.0026	0.0016
8		0.0006		0.0008		0.0004
9 -				0.0004		0.0002
no. of sample words	1035	1678				
$\sum \lvert p_o - p_e \rvert$			0.0228	0.1640	0.0886	0.1060
Kolmogorov distance (%)			0.65	5.08	2.01	2.72

1/ For explanation of LN(a), LN(b), and θ and λ vide footnote to Table 6.3.
2/ $\theta = 0.1624$, $\lambda = 0.1984$;
3/ $\theta = 0.2487$, $\lambda = 0.1961$
4/ $\theta = 0.2965$, $\lambda = 0.1588$
5/ $\theta = 0.3657$, $\lambda = 0.1570$.

Chapter 7: A classification of Bengali syllables

7.1.1. Introductory: In the present chapter, we introduce a simple classification of Bengali syllables, and study the relative frequencies of the different classes in Bengali prose and poetry texts. This work was undertaken, originally, for improving upon the number of syllables as a measure of word-length.

7.1.2. Bengali syllables rarely contain more than four phonemes; the syllable 'strai' is quite exceptional. English syllables show much greater variation in respect of length; thus, the word 'strength' is monosyllabic. But even though Bengali syllables do not vary greatly in length, they do vary sufficiently to make the number of syllables an inadequate measure of word-length. So attempts were made to improve upon this measure.

7.1.3. Bengali syllables seem to fall into two more or less homogeneous types, to be called types A and B, if one is interested in their length or in the time normally required to pronounce them. The definitions are given below:

type		definition[1]	illustration
A: short		open i.e., vowel-ending syllables without diphthongs	o, ma, kha, sra
B: long	B_1	closed i.e., consonant-ending syllables with/without diphthongs	an, nun, snan, bang, aik
	B_2	open syllables with diphthongs	ai, mao, strai

1/ For definitions of open and closed syllables, vide S.K.Chatterjee (1945, pp. 25, 35).

It will be seen that type B syllables have been further subdivided into sub-types B_1 and B_2. But this distinction was not made in all the counts, and is less important than that between types A and B, at least for the present purposes.

7.1.4. Generally speaking, type B syllables are longer than type A syllables, and, and and, as already stated, the two types are more or less homogeneous in length. For purposes of metric analysis, type B syllables are sometimes supposed to take two mora or instants for pronunciation, as against one mora required by type A syllables (Chatterjee, 1945, pp. 377-8). In any case, instead of stating the average word-length for Bengali prose work as 2.1 syllables per word, say, it would be preferable to say that the average word has 1.4 syllables of type A and 0.7 syllables of type B (say).

7.1.5. For comparisons between works in a given language, such refinements may not be important at all, for the relative frequencies of the different types of syllables may be more or less stable within a language [vide Table 7.1 below]. The position is different when one tries to compare average word-lengths for different languages. Compare English and Bengali, for example. At first sight, there appears to be a wide gap between the two averages, viz., about 2.25 syllables per word for Bengali prose and about 1.45 syllables per word for English prose. This difference appears to be much less if one is told that in Bengali prose about two-thirds of the syllables belong to type A ('short') and only one-third come under type B ('long'), while in English prose the relative frequencies seem to be practically reversed,

7.1.6. This approach can give only rough indications. English syllables, for example, cannot be split into two homogeneous types; a larger number would be necessary. And ultimately, one must use some conversion factors for expressing different types of syllables in terms of a standard type or in terms of a common unit like mora. Otherwise one cannot compare the two languages in a conclusive manner. But the choice of conversion factors will have to be partly subjective and therefore open to criticism. It is obviously simpler and better to measure word-length in terms of letters and phonemes (vide Appendices 1 and 3 for studies on word-length in letters).

7.1.7. This study on relative frequencies of different types of syllables may therefore be taken more as a study on the structure of Bengali syllables than as an investigation concerned with word-length.[1/] Some data are also presented on the relative frequencies of different types of syllables in different positions within words of different lengths. The relative frequencies of different types of syllables seem to be fairly stable in Bengali prose. But the percentages vary to a much greater extent in Bengali poetry, and this variation seems to be related to the meter used in the poems.

7.2.1 Relative frequencies of different types of syllables in Bengali prose : Table 7.1 shows the estimated percentages of syllables belonging to type A, separately for sixteen works, mostly fiction, three short essays and three short stories. All these were covered in word-length studies (vide Chapter 3), and the same samples of words

---

1/ For some earlier studies on structure of syllables, vide Chapter 1, Section 1.3.

were used here, excepting the systematic samples from certain works subjected to both methods of sampling. The estimated percentages are given for the four sub-samples comprising the sample; the 'combined' estimates are also given. Table 7.2 presents similar percentages for type B_2 syllables for a subset of works covered in Table 7.1.

7.2.2. All these estimates have the ratio form and excepting perhaps for the estimates for type B_2, the estimates may safely be assumed to possess the large sample properties of ratio estimates (vide Chapter 2, Sections 2.4-5).

7.2.3. The first point to note is the broad consistency in the percentages of type A syllables for the different works. The combined percentages in Table 7.1 are spread over the range 61.88-71.61, the average being nearly 67. There are no doubt appreciable differences between works, and even within authors, and some of these are found to be significant. (The subsample estimates indicate that the standard errors of the combined percentages are usually less than 1%.) But it is very difficult to interpret these differences. The percentage of type A syllables may vary between different classes of words like 'tatsama' (Sanskrit), 'tatbhava' (Prakrit), indigenous (non-Aryan) and foreign (loan) words, and also between conversational and other words. It is possible that the between works differences in relative frequencies of these classes of words produced the differences in relative frequencies of syllable-types.

Table 7.1 Estimated percentages of 'short' or type A syllables[1/] separately for different works in Bengali prose and by subsamples.

author	work	type of sample	no.of sample words	total no. of syllables by subsamples					percentage of type A (short) syllables				
				ss1	ss2	ss3	ss4	comb.	ss1	ss2	ss3	ss4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Vidyasagar	Shakuntala	prob.	696	483	448	470	481	1882	68.32	67.41	71.49	69.03	69.08
	Sitar Vanavas	"	750	508	509	499	505	2021	66.34	67.19	67.34	65.35	66.55
Bamkimchandra	Durgeshnandini	"	577	397	362	363	366	1488	67.76	64.64	66.39	68.31	66.80
	Visavriksha	"	611	363	370	372	404	1509	71.07	68.11	69.89	68.32	69.32
Rabindranath	Gora	"	889	488	520	522	542	2072	70.90	69.42	70.88	66.79	69.45
	Sheser Kavita	"	735	399	415	402	391	1607	65.41	67.95	68.41	68.03	67.45
Iramatha Chaudhuri	Char-Yari Katha	"	872	459	429	447	461	1796	66.67	65.03	66.22	64.43	65.59
	Birbaler Halkhata	"	1041	623	617	560	606	2406	61.80	61.75	60.54	63.37	61.88
Saratchandra	Pallisamaj	"	890	485	495	487	502	1969	72.37	70.71	70.02	69.32	70.59
	Pather Dabi	"	815	504	453	441	418	1816	70.44	69.76	69.39	68.18	69.49
Bibhutibhusan	Pather Panchali	"	922	507	552	514	502	2075	73.57	71.20	69.26	69.72	70.94
	Devayan	"	931	506	505	492	476	1979	67.59	66.73	70.73	71.43	69.08
Jajabar	Dristipat	"	772	450	468	471	455	1844	62.44	65.81	64.54	60.22	63.29
	Janantik	"	690	386	384	403	409	1582	64.77	62.24	67.74	60.15	63.72
Muztaba Ali	Chacha-Kahini	"	778	420	458	408	417	1703	66.67	65.72	65.20	63.79	65.36
	Deshe Videshe	"	791	445	403	428	442	1718	68.76	64.02	62.85	63.80	64.90
Bamkimchandra	Samya	syst.	1010	668	659	685	633	2645	66.02	66.01	66.71	64.30	65.79
Rabindranath	Bamkimchandra	"	1237	880	800	803	835	3318	63.64	63.00	62.14	62.40	62.81
	Vishwavidyalay	"	1009	594	579	590	597	2360	66.83	63.90	64.74	62.14	64.41
	Kabuliwala	"	779	483	488	485	492	1948	68.74	73.98	72.17	71.55	71.61
	Kshudhita Pasan	"	1192	754	749	744	761	3008	66.71	68.36	70.03	70.83	68.98
	Laboratory	"	1228	697	596	679	645	2617	74.60	68.45	69.36	69.45	70.58

1/ that is, open or vowel-ending syllables without diphthongs

Table 7.2: Estimated percentages of type B_2 syllables[1/] for selected works in Bengali prose and by subsamples.

author	work	type of sample	no.of sample words	total no. of syllables by subsamples					percentage of type B_2 syllables				
				ss1	ss2	ss3	ss4	comb.	ss1	ss2	ss3	ss4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Vidyasagar	Shakuntala	prob.	696	483	448	470	481	1882	4.97	5.13	4.04	3.74	4.46
	Sitar Vanavas	"	750	508	509	499	505	2021	5.51	4.52	5.41	3.56	4.75
Saratchandra	Pallisamaj	"	890	485	495	487	502	1969	6.39	7.47	4.72	6.77	6.35
Pramatha Chaudhuri	Birbaler Halkhata	"	1041	623	617	560	606	2406	6.58	5.02	7.50	6.27	6.32
Jajabar	Janantik	"	690	386	384	403	409	1582	5.18	4.95	2.48	5.62	4.55
Bamkimchandra	Samya	syst.	1010	668	659	685	633	2645	5.09	4.70	6.40	3.79	5.02
Rabindranath	Bamkimchandra	"	1237	880	800	803	835	3318	5.00	5.00	6.23	4.91	5.27
	Vishwavidyalay	"	1009	594	579	590	597	2360	4.38	5.01	3.90	4.86	4.53
	Kabuliwala	"	779	483	488	485	492	1948	5.38	4.30	4.74	5.69	5.03
	Kshudhita Pasan	"	1192	754	749	744	761	3008	5.04	4.54	3.63	3.81	4.26
	Laboratory	"	1228	697	596	679	645	2617	3.59	4.87	5.60	4.81	4.70

1/ that is, (long) syllables ending in diphthongs.

7.2.4. There is little to comment on the percentages of the type B_2 syllables shown in Table 7.2, except that the percentages are all of the order of 5%.

7.3.1. Relative frequencies of syllable-types in Bengali poetry: Table 7.3 presents some similar data for Bengali poetry. The material is the same as that studied in Chapter 3. [Vide Chapter 2, Section 2.8, where the division of poems into parts is explained.] One finds considerable variation in the percentage of type A syllables for different poems/poetry pieces. This makes a sharp contrast with the findings for Bengali prose. Whereas in Bengale prose, the percentage of type A syllables varies within a narrow range, from about 62% to 72% , the percentage is seen to vary from 65 to 90 among the different pieces of poetry examined. Also such variation seems to be related to the meter of the poem. It is not safe to generalise from a small sample of poems, so the observations in the following para should be regarded as tentative.

7.3.2. The three poems in free verse, viz., "Banshi", "Ami" and "Africa", show percentages lying between 65 and 72. In this respect, therefore, free verse is completely similar to the prose works covered in Table 7.1. One poem for children, viz., "Khelabhola," employs an accent-dominated meter, and the percentage is nearly 70. The highest percentage, about 90, is shown by the lyric, "Varsamangal", which uses the 'mattravritta' meter. All other poems are in the traditional "Payar" and related meters, where the vocal drawl is predominant. The percentage of type A syllables varies widely within this category

from about 70 (cf. "Sandhya") to about 87 (cf. "Sonar Tari"). Thus, the extract from "Meghanadabadha Kavya" employs the blank verse variant of "Payar" and shows a percentage near 80. To a certain extent, perhaps, poems in an elevated or serious tone have lower percentages than poems with a pronounced lyrical quality. Again, it is possible that poems employing unequal lines, but not stanzas (e.g., "Shajahan", "Balaka" and "Ora Kaj Kare") tend to have lower percentages, and resemble the prose poems already discussed.[1/]

7.3.3. On the whole, Bengali poetry has progressed towards the free verse and away from the traditional "Payar" and similar meters. So, in all probability, there has been a declining trend in the percentage of type A syllables, and hence an upward trend in the percentage of type B syllables, in Bengali poetry over the last century.

7.3.4. As already stated (vide Chapter 3, Section 3.3) word-length does not have the same significance in Bengali poetry as in Bengali prose. Nevertheless, variation in the percentage of type A syllables vitiates between poems comparisons of word-length in syllables.

7.3.5. There is little to say about the figures for type B_2 syllables in Table 7.3. One may note, however, that the percentages are about 3, on the average, which is smaller than in Bengali prose, where the percentage is nearly 5, on an average. Also, the percentages in Table 7.3 fluctuate more widely than those in Table 7.2, but this may be partly due to the smaller sample sizes in Table 7.3.

1/ A lyric by Satyendranath Datta, entitled "Sindhutandav" has not been analysed here, but the meter used /imitative of "Pancha-chamar" meter of Sanskrit poetry/ requires that syllables of types A and B must alternate, so that each type forms exactly 50% of the total.

Table 7.3: Percentages of different types of syllables in Bengali poetry, separately for different poems/poetry pieces and by parts[1]

author	work	piece/poem	year of publication/ composition	no. of words	no. of syllables by parts			percentage of type A_2[2] syllables			percentage of type B_2[3] syllables		
					pt.1	pt.2	comb.	pt.1	pt.2	comb.	pt.1	pt.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Michael M. Datta	Meghanadabadha Kavya	First 200 lines of Canto I	1861	981	1336	1188	2524	80.3	79.1	79.8	2.3	2.0	2.2
Rabindranath	Prabhat Sangit	Nirjharer Swapnabhanga	1882	221	254	212	466	78.7	80.2	79.4	3.2	2.8	3.0
	Manasi	Badhu	1888	409	451	420	871	86.9	83.3	85.2	2.2	3.8	3.0
		Meghadut	1890	594	677	1014	1691	73.1	76.8	75.4	1.3	2.5	2.0
	Sonar Tari	Sonar Tari	1892	171	188	174	362	91.5	82.8	87.3	3.2	6.9	5.0
		Puraskar	1894	314	366	412	778	79.8	82.5	81.2	2.2	3.6	3.0
		Niruddesh Yatra	1893	292	343	344	687	83.7	84.0	83.9	2.3	2.9	2.6
	Chitra	Sandhya	1894	279	296	390	686	66.9	72.8	70.2	4.0	1.8	2.8
		Urvashi	1895	325	465	464	929	74.0	78.4	76.2	1.7	1.1	1.4
	Kalpana	Varsamangal	1897	206	295	395	690	90.2	90.4	90.3	3.0	1.3	2.0
		Swapna	1897	236	363	272	635	72.2	80.5	75.7	1.6	1.5	1.6
	Kshanika	Krismakali	1900	190	152	230	382	75.0	73.9	74.3	4.0	4.8	4.5
	Shishu	Virpurus	1903	326	318	318	636	78.0	78.6	78.3	5.7	3.1	4.4
	Balaka	Shajahan	1914	575	869	504	1373	70.3	76.8	72.7	5.9	4.4	5.3
		Balaka	1915	268	258	405	663	76.0	73.1	74.2	1.9	3.7	3.0

1/ All poems/pieces except "Puraskar" were completely counted. For "Puraskar", a systematic sample of stanzas was used.

2/ that is, open or vowel-ending syllables without diphthongs.

3/ that is, open syllables ending in diphthongs.

(contd.)

Table 7.3 (Contd.) Percentages of different types of syllables in Bengali poetry, separately for different poems/poetry pieces and by parts[1]

author	work	piece/poem	year of publication/ composition	no. of words	no. of syllables by parts			percentage of type A[2] syllables			percentage of type B[2] syllables[3]		
					pt.1	pt.2	comb.	pt.1	pt.2	comb.	pt.1	pt.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Rabindranath	Shishu Bholanath	Khelabhola	1921	220	219	221	440	68.9	72.4	70.7	5.5	5.9	5.7
	Purabi	Satyendranath Datta	1922	637	819	759	1578	72.4	74.8	73.5	2.7	3.0	2.9
		Tapobhanga	1923	557	731	746	1477	68.8	73.3	71.1	1.9	2.1	2.0
	Parishes	Pranam	1931	244	291	383	674	70.8	75.2	73.3	2.1	1.6	1.8
		Banshi	1932	312	306	419	725	71.2	72.1	71.7	4.9	5.0	5.0
	Shyamali	Ami	1936	265	265	314	579	70.6	65.6	67.9	2.6	2.6	2.6
	Samayik Patra	Africa	1937	214	213	334	547	67.1	64.1	65.3	2.4	3.3	2.9
	Arogya	Ora Kaj Kare	1941	199	232	260	492	72.8	77.7	75.4	4.7	3.8	2.4

1/ All poems/pieces except "Puraskar" were completely counted. For "Puraskar", a systematic sample of stanzas was used.

2/ that is, open or vowel-ending syllables without diphthongs.

3/ that is, open syllables ending in diphthongs.

7.4.1. Composition of Words in terms of different types of syllables: Table 7.4 shows the distribution of words coming in the probability samples from all the sixteen works first listed in Table 7.1 taken together, according to the number of syllables of type A and type B comprising the word. The figures at the foot of the table are of the greatest interest.

7.4.2. Some data are presented in Tables 7.5 and 7.6 on the composition of words in terms of different types of syllables, or concretely, on the relative frequencies of syllable-types in different positions within words of different lengths. Table 7.5 is based on the probability samples of words from five selected works taken together. Each sample word was first given a composition code. The code '$B_1 A B_2$', for instance, was given to a three-syllabled word, where the first syllable was of type B_1, the second of type A and the third of type B_2. Table 7.5 is simply the frequency distribution of all words in the probability samples from the five works taken together, according to such composition codes.

Table 7.4: Percentage distribution of words by numbers of short (type A) and long (type B) syllables, based on probability samples from sixteen works[1] combined[2].

total no. of sample words: 12760

no. of type A syllables	no. of type B syllables						
	0	1	2	3	4	5	total
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	-	11.00	5.12	0.28	0.016	-	16.41
1	7.05	23.12	3.98	0.40	0.05	-	34.60
2	19.25	10.08	1.83	0.35	0.08	0.08	31.53
3	9.71	3.03	0.77	0.09		0.016	13.61
4	2.02	0.79	0.11	0.06			2.99
5	0.49	0.20	0.04			0.08	0.74
6	0.04	0.02	0.016				0.076
7	0.02	-	0.028				0.028
10	=	0.028	-				0.008
total	38.59	48.25	11.87	1.19	0.07	0.03	100.00

1/ The first sixteen works listed in Table 7.1.
2/ Average no. of syllables per word : type A – 1.55, type B – 0.76.
Percentage of all syllables : type A-67.09, type B - 32.91.

Table 7.5 : Frequency distribution of words according to composition codes*, based on probability samples of words from Shakuntala, Sitar Vanavas, Birbaler Halkhata, Pallisamaj and Janantik, combined.

composition code* (1)	no.of words (2)	composition code* (1)	no.of words (2)	composition code* (1)	no.of words (2)
A	287	$B_1B_1B_2$	1	B_1B_1AA	11
B_1	244	B_1B_2A	1	$B_1B_1AB_1$	4
B_2	173	B_2AA	17	$B_1B_1AB_2$	1
s.t.	704	B_2AB_1	11	$B_1B_1B_1A$	1
AA	645	B_2B_1A	3	$B_1B_1B_1B_1$	1
AB_1	425	$B_2B_1B_1$	1	$B_1B_1B_1B_2$	1
AB_2	74	s.t.	1108	B_1B_2AA	2
B_1A	305			$B_1B_2AB_1$	1
B_1B_1	138	AAAA	90	B_2AAA	3
B_1B_2	27	$AAAB_1$	47	B_2AAB_1	2
B_2A	83	$AAAB_2$	2	B_2AB_1A	2
B_2B_1	21	AAB_1A	15	B_2B_1AA	1
B_2B_2	7	AAB_1B_1	15	s.t.	341
s.t.	1725	AAB_2A	3	AAAAA	25
AAA	406	AB_1AA	29	$AAAAB_1$	8
AAB_1	255	AB_1AB_1	22	$AAAB_1A$	3
AAB_2	30	AB_1AB_2	2	$AAAB_1B_1$	3
AB_1A	93	AB_1B_1A	14	AAB_1AA	4
AB_1B_1	54	$AB_1B_1B_1$	1	AAB_1AB_1	2
AB_1B_2	8	$AB_1B_1B_2$	2	AAB_1B_1A	8
AB_2A	11	AB_2B_1A	1	$AAB_1B_1B_1$	1
AB_2B_1	1	$AB_2B_1B_2$	1	AAB_2B_1A	1
AB_2B_2	1	B_1AAA	35	AB_1AAA	15
B_1AA	79	B_1AAB_1	7	AB_1AAB_1	3
B_1AB_1	68	B_1AAB_2	3	AB_1AAB_2	1
B_1AB_2	6	B_1AB_1A	12	AB_1AB_1A	6
B_1B_1A	51	$B_1AB_1B_1$	8	$AB_1AB_1B_1$	2
$B_1B_1B_1$	11	$B_1AB_1B_2$	2	AB_1B_1AA	2

* For definitions of different types of syllables, see para 7.1.3; for definition of composition codes see para 7.4.2.

(contd.)

Table 7.5 : (Contd.)

composition code* (1)	no.of words (2)
$AB_1B_1AB_2$	1
$AB_1B_1B_1A$	1
$AB_1B_1B_1B_1$	1
AB_2AB_1A	1
B_1AAAA	11
B_1AAAB_1	5
B_1AAAB_2	2
B_1AAB_1A	4
$B_1AAB_1B_1$	5
B_1AB_1AA	4
$B_1AB_1AB_1$	4
$B_1AB_1B_1A$	2
$B_1AB_2B_1A$	2
B_1B_1AAA	3
$B_1B_1AAB_2$	1
$B_1B_1AB_1B_1$	1
$B_1B_1B_1AA$	2
$B_1B_1B_1B_1A$	1
B_2AAB_1A	1
$B_2AAB_1B_1$	1
s. t.	137
$AAAAAA$	4
$AAAAAB_1$	1
$AAAAAB_2$	1
$AAAB_1AA$	1
AAB_1AAA	5
AAB_1AAB_1	1
AAB_1AAB_2	1
$AAB_1AB_1B_1$	2
AAB_1B_1AA	1
$AAB_2B_1B_1A$	1
AB_1AAAA	4
AB_1AAB_1A	1
AB_1AB_1AA	2
$AB_1AB_1AB_1$	1
$AB_1B_1AAB_1$	1
B_1AAAAA	2
B_1AAAAB_1	2
$B_1AAAB_1B_1$	1
$B_1AAB_1AB_1$	2
B_1B_1AAAA	1
$B_1AAB_1B_1A$	2
$B_1AB_1B_1AA$	1
s. t.	38
$AAAAAAA$	2
$AAAB_1B_1AB_1$	1
AB_1AAAAA	1
$AB_1AAAB_1B_1$	1
$B_1AAAAAB_1$	1
$B_1AABAAB_1$	1
$B_1AB_1AAAB_1$	1
$B_1AB_1B_1AAA$	1
$B_1AB_1B_1AB_1B_1$	1
s. t.	10
B_1AB_1AAAAA	1
$B_2AB_1AB_1B_1B_1A$	1
s. t.	2
B_1AAAB_1AAAA	1
$B_1B_1AB_1B_1AAAAB_1$	1
s. t.	2
grand total	4067

* For definitions of different types of syllables, see para 7.1.3.; for definition of composition codes see para 7.4.2.

7.4.3. • Ignoring the distinction between sub-types B_1 and B_2 and also ignoring the order of syllables within words, Table 7.5 can be condensed into one analogous to Table 7.4 and the latter can in turn give an ordinary one-way word-length distribution like those given in Chapter 3.

7.4.4. Table 7.6 is derived from Table 7.5 and gives the percentages of different types of syllables in specified positions of words of specified length. Thus, there are 1108 three-syllabled words in the sample, and hence 1108 final syllables of three-syllabled words. The table shows that 59.66% of these are of type A, 36.19% of type B_1 and only 4.15% of type B_2.

Table 7.6 : Percentages of different types of syllables by word-length and by position of syllable within word, based on probability samples of words from Shakuntala, Sitar Vanavas, Birbaler Halkhata, Pallisamaj and Janantik, combined.

word-length in syllables	no.of sample words	position of syllable within word	no.of sample syllables	percentage of syllables by types		
				A	B_1	B_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	704	1	704	40.77	34.66	24.57
2	1725	1	1725	66.32	27.25	6.43
		2	1725	59.88	33.86	6.26
		all	3450	63.10	30.55	6.35
3	1108	1	1108	77.53	19.58	2.89
		2	1108	78.70	20.04	1.26
		3	1108	59.66	36.19	4.15
		all	3324	71.96	25.27	2.77
4	341	1	341	71.55	26.10	2.35
		2	341	72.14	26.39	1.47
		3	341	76.83	22.29	0.88
		4	341	64.22	31.67	4.11
		all	1364	71.19	26.61	2.20

(contd.)

Table 7.6: (Contd.)

word-length in syllables	no. of sample words	position of syllable within word	no. of sample syllables	percentage of syllables by types		
				A	B_1	B_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
5	137	1	137	64.23	34.31	1.46
		2	137	70.07	29.20	0.73
		3	137	73.72	24.09	2.19
		4	137	67.88	32.12	-
		5	137	70.07	26.28	3.65
		all	685	69.19	29.20	1.61
6	38	all	228	71.49	27.19	1.32
7	10	all	70	67.14	32.86	-
8 & above	4	all	35	60.00	37.14	2.86
all	-	1	4067	65.26	26.70	8.04
	-	2	3363	67.98	28.22	3.81
	-	3	1638	64.59	32.17	3.24
	-	4	530	65.66	31.70	2.64
	-	5	189	72.49	24.87	2.65
	-	6	52	69.23	26.92	3.85
	-	7 & above	21	61.90	38.10	-
all	4067	all	9860	66.25	28.39	5.37

7.4.5. Certain interesting points emerge from Table 7.6. The most striking feature is the higher percentage for type B_2(24.57%) and the lower one for type A (40.77%), among monosyllabic words. This is seen to a smaller extent for bi-syllable words also. Percentages for type B_2 are generally low for words with more than two syllables. Thus, type B_2 syllables are mostly found in monosyllabic or bi-syllable

words. Similarly, percentages for type A are higher for words with three or more syllables.

7.4.6. It also appears that for words with 2, 3 or 4 syllables the proportion of type A syllables is comparatively low among final syllables of words than in other positions; and the percentages of type B syllables are correspondingly higher.[1] Such differences are not clear for words with 5 syllables or more, where the sample sizes are also rather small. But if we are interested in type B_2 syllables, the percentage is higher in the final position than in other positions for words with 3, 4 or 5 syllables.

7.4.7. Finally, the percentage of type A syllables in the initial position seems to vary to some extent between words having 2, 3 or 4 syllables.

7.4.8. The differences mentioned in paras 7.4.5 - 7.4.7 are generally small, so the relative frequencies of syllable-types are more or less the same in different positions of words of different lengths. The only notable exception is seen for syllables forming monosyllabic words, where the distribution is very different.

7.4.9. Such tables could be examined separately by works and also by sub-samples. Indeed, a preliminary examination showed that the five works give appreciably different tables. The differences mentioned above are very marked for some works, and less marked for others.

1/ Many of the final type B_1 syllables owe their origin to the gradual dropping of the final vowel sound /vide S.K.Chatterjee, 1945, pp. 28-29/. As against this, type B_1 syllables in other positions are generally associated with conjunct consonants.

7.4.10. At the instance of Dr. K. R. Parthasarathi of the Indian Statistical Institute, an attempt was made to fit simple Markov chains to sample texts considered as sequences of four types of elements, viz., the three types of syllables, A, B_1 and B_2, and the gap g between words. Data of Table 7.5 were used for the purpose. (The heterogeneity has not probably affected the conclusion.) It was found that the simple Markov chain could not be an adequate model for the language. The position did not improve when the distinction between sub-types B_1 and B_2 was dropped. Table 7.7 brings out the reason : The preceding and the following states are far from statistically independent. No attempt was made to fit more general stochastic processes.

Table 7.7: Joint distribution of states preceding and following specified states, based on data of Table 7.5.

(a) specified state : type B_2 syllable

preceding state	following state				total
	g (gap)	A	B_1	B_2	
g (gap)	173	121	26	7	327
A	125	15	7	1	148
B_1	42	4			46
B_2	8				8
total	348	140	33	8	529

(b) specified state : type B syllable

preceding state	following state			total
	g (gap)	A	B	
g (gap)	417	702	294	1413
A	1002	277	155	1434
B	320	129	32	481
total	1739	1108	481	3328

Chapter 8 : Sentence-length in Bengali Prose - (1)

8.1.1. Introduction : In this and the following chapter are reported some statistical studies on sentence-length almost entirely confined to Bengali prose.[1]

8.1.2. Sentence-length is, like word-length, one of the most obvious and most important indicators of style. It has often been said that [Williams, 1940, p.356; Herdan, 1964, Chap. 1, Section 1.3] sentence-length is likely to give a more sensitive measure of style than word-length, being less language-conditioned than the latter.[2]

8.1.3. Sentence-length is one of the two components of the Reading Ease Score in the wellknown Readability Scale due to Rudolph Flesch (1948). The other component is the number of syllables per 100 words (vide Chapter 1, Section 1.2).

8.1.4. A pioneering study on sentence-length in English and Latin prose was made by Yule (1938) who used sentence-length data for solving two cases of disputed authorship. [Vide Chapter 1, Section 1.2]. Yule measured sentence-length in terms of the number of words. The same practice has been adopted in the present study. Many other studies have been reviewed by Miller (1951, Chapter 6) — some of these are earlier than Yule's — where one may find a dimensional idea of the average

---

1/ Some data on the English novel, "Pride and Prejudice", are used in Chapter 9.

2/ The third important indicator is the vocabulary, its size and composition, [vide Chapter 1, Section 1.2]. It is felt that for Bengali, a very important style-indicator would be the percentage of distinct words or word-occurrences belonging to different etymological groups like 'tatsama' (Sanskrit) and 'tadbhava' (Prakrit). [Vide Chap. 1, Section 1.5, in this connection.]

length of sentences in different types of writing. In one study mentioned by Miller sentence-length was measured in syllables instead of words. This is no doubt an improvement. But if word-length and sentence-length studies are both carried out on the same work, one can multiply the average sentence-length in words by the average word-length in syllables to estimate the average sentence-length in syllables. And it does not seem to be happy to mix up the two indicators of style into one single measure.

8.1.5. Flesch (1946, 1948) considers the effect of variation in sentence-length on readability and indicates sentence-length averages for different types of writing (levels of readability) in English prose (vide Chap. 1, Section 1.2).

8.1.6. Williams (1940) showed that the distribution of sentence-length in terms of words is approximately lognormal, for three books on sociological topics, one by each of three English authors (Chesterton, Shaw and Wells). Williams did not use any goodness of fit measure and depended on a visual examination of some elementary graphs. The mean and the s.d. of the underlying logarithmic variate should therefore suffice to specify the distribution. Williams found that the three English authors not only varied in respect of the mean of the logarithmic variate — which indicates location or central tendency — but they also varied in respect of the s.d. of the logarithmic variate, which controls the coefficient of variation.[1/2/]

1/ If x is lognormally distributed with the expectation $E(\log_e x) = \theta$ and the variance $V(\log_e x) = \lambda^2$, then the median or geometric mean of x is $e^{\theta} = e^{E(\log_e x)}$ and C.V. of x is $\sqrt{e^{\lambda^2} - 1}$, that is $\sqrt{e^{V(\log_e x)} - 1}$.

2/ Why sentence-length should be lognormally distributed is difficult to answer with any degree of confidence. [Vide, however, Williams (1940, pp.360-361).]

8.1.7. Subba Rao (1960) carried out a study on sentence-length in eight works in Kannada prose written by three different authors. From each work he selected about 50 pages by probability sampling (of the systematic type) and counted lengths of all sentences appearing on the sample pages. Between author differences were shown to be significant, but within author differences were not shown to be nonsignificant or negligible, which should have been done before concluding that sentence-length is an indicator of an author's style. Subba Rao tacitly assumed that the distributions are lognormal, without making any examination at all. Williams' finding that sentence-length distributions are lognormal rests on a small volume of material relating to essays in English. It is necessary to supplement Williams' study by others based on diverse types of material. Subba Rao also applied statistical tests valid when unrestricted random sampling is used, without any investigation of the randomness of the series of sentence-lengths. The same assumption was implicitly made by Williams (1940, p. 357) in calculating standard errors of estimates of parameters. Both had used samples of sentences with a great deal of clustering involved in them.

8.1.8. Studies on sentence-length reported in this and the next chapter are roughly parallel to those on word-length described in Chapters 2 to 6. But the investigations on sentence-length are on a smaller scale, and poetry has been left out, for obvious reasons. In this chapter we discuss the following:

(i) the method of probability sampling used for the sampling of sentences;

(ii) the sentence-length distributions thrown up for Bengali prose, which give dimensional ideas about sentence-length in different types of texts;

(iii) the between and within author differences which throw light on the efficacy of sentence-length as a statistical indicator of an author's individuality; and

(iv) the form of the sentence-length distributions for Bengali prose, vis-a-vis the lognormal hypothesis put forward by Williams (1940).

8.1.9. It will be seen that in Bengali prose, the sentence-length distribution may vary significantly and appreciably between similar works by the same author, so that sentence-length is not a definite indicator of an author's style. It will also be seen that the lognormal distribution gives a fairly good fit to the sentence-length distributions for Bengali prose.

8.2.1. Definition of Sentence-length : A statistical study of sentence-length requires a more precise definition of sentence than is commonly given. If the length is defined in terms of the number of words, the problem of defining the word also arises.

8.2.2. These difficulties were discussed at length by Yule (1938, especially pp.363- 5) who was considering sentence-length in English and Latin prose. Yule concluded that the punctuation was unreliable — he was considering works of nineteenth and earlier centuries when punctuation did not get the same attention as now and compositors had much say in the matter. Yule revised the punctuation before measuring sentence-length.

8.2.3. In the present case, both the problems were not quite serious, generally speaking. Words were taken as printed, without any attempt to treat compounds in any special manner /vide Chapter 2, Section 2.2/. Also, the punctuation as found was generally accepted, excepting in a very small proportion of cases; for "Pather Panchali", in particular, some very inappropriately used dashes were treated as termination marks.

8.2.4. One point deserves to be emphasised here. According to many workers the sentence should be defined as a single unit-of-thought, and semicolons and colons should not be taken as termination marks of sentences. The other view is that semicolons and colons demarcate sentences /Flesch (1946), Chap. 4; see also p.93 and p.195/. The first definition of sentence was used in the present study, and semicolons and colons were not regarded as termination marks, excepting where a colon introduces a speech with more than one sentence. Semicolons are very frequent in Vidyasagar's works and in many cases the full-stop should have been used instead. This point will be discussed in Section 8.5 below. Excepting for Vidyasagar's works, the main findings would not be greatly altered if the other definition of sentence were used.

8.3.1. Probability sampling of sentences : Sentence-length studies carried out so far have generally been based on non-probabilistic samples, if not on complete enumeration. Yule (1938) mostly used what he called "the method of selected passages of considerable length" (Yule, 1938, p.383) selecting a number of passages evenly spread over the work or works sampled. He considered this method superior to probability

sampling on the ground that non-representative matter may be deliberately excluded from the former type of sample. But this claim is not quite correct, for a probability sample is also capable of adaptation.

8.3.2. Yule discussed the difficulties of taking probability samples of sentences (ibid, pp. 381- 3), but did not seem to find any convenient method of probability sampling. He actually made some crude attempts at probability sampling (ibid, pp. 374-5), selecting columns at random from Gerson's works and counting 20 sentences from each selected column, or pages at random from Petty's, taking 10 sentences from each selected page. Yule also used a rough analogue of Mahalanobis' technique of interpenetrating subsamples : he divided the set of passages selected from a given work into two subsets and used the divergence between the two distributions as indicator of sampling error.

8.3.3. Williams (1940) selected the first 30 sentences from each of first 20 chapters from the work by Chesterton; the greater part of the work was covered in this manner. Similar procedureswere adopted for the works by Wells and Shaw. Flesch (1946, Appendix, 1948) recommends following strictly numerical rules for selecting samples, which means selecting some kind of a systematic sample.

8.3.4. Much of the material utilised in the present study is based on probability samples; systematic samples were taken from a few works, and their use justified by statistical tests. The measurement of sampling error has been done by the technique of IPNS. Some texts or extracts were also subjected to complete counts.[1/]

1/ Probability sampling of pages was used by Subba Rao (1960), but methods suitable for unrestricted random sampling were used in an uncritical manner (see para 8.1.7 supra).

8.3.5. For nineteen of the twenty works listed in Table 8.2 probability samples of sentences were selected. From three of these nineteen works, "systematic" samples of sentences were also selected in the manner described below. For the remaining work, viz., "Chaturanga", the "sample" was that obtained by complete enumeration of the first two of the four parts of the work.[1/]

8.3.6. It may be recalled that for drawing probability samples for word-length studies, 100 or 200 (say) lines were selected at random, with equal probability and with replacement, from the work under study; and all words falling on all the sample lines together formed the probability sample of words. [Vide Section 2.3 of Chapter 2 for details]. These randomly selected lines were again utilised for probability sampling of sentences; but for some of the 19 works, the number of sample lines taken for word-length studies seemed to be rather small for sentence sampling purposes, and the sample of lines was extended by following the same type of procedure. For "Kavi Shri Ramakrishna", of course, this selection of lines had to be done anew for sentence-length studies.

8.3.7. Once a random sample of lines from the work under study is available, one can pick out the sentences having termination marks or, in plain words, ending on these sample lines. Some sample lines do not give any such sentence, while others give one, two or even more. All such sentences taken together would constitute a probability sample of sentences from the work.

1/ Only one out of these 20 works, viz., "Kavi Shri Ramakrishna" by Achintya Sengupta was not covered in the word-length studies also.

8.3.8. For the different works, 100 sample lines generally gave less than 100 sample sentences. Typically, about 60 to 70 sentences were obtained. So the procedure was modified to get a larger sample of sentences. Instead of taking only the sentence(s) having termination mark(s) on the sample line, we also included the one or more sentences ending on the line just preceding the sample line which had at least one termination mark, and also the one or more sentences ending on the line just following the sample line which had at least one termination mark. But if the sample line itself did not give any sample sentence; no such extension was done. This virtually tripled the sample size at the cost of introducing some further clustering of sample sentences. Without this extension the probability sample of sentences would be almost random, for most of the sample lines do not give more than one sentence ending on them.

8.3.9. It is necessary to examine the nature of the probability sample of sentences drawn in the above-mentioned manner. Suppose one lists all the lines in a work having one or more termination marks, gives a serial number to these lines following the natural reading order and writes against each serial number the lengths of sentences ending on the corresponding line. One then gets a scheme like the the following. Frequently n_i's are 1; n_i=2 in a small proportion of cases, but n_i = 3 or more seems to be very rare.

sr. no. of line having at least one termination mark	lengths of sentences ending on each line
(1)	(2)
1	$l_{11}, l_{12}, \ldots\ldots, l_{1n_1}$
2	$l_{21}, l_{22}, \ldots\ldots, l_{2n_2}$
.	
.	
.	
i -1	$l_{i-1,1}, l_{i-1,2}, \ldots\ldots, l_{i-1,n_{i-1}}$
i	$l_{i1}, l_{i2}, \ldots\ldots\ldots, l_{in_i}$
i + 1	$l_{i+1,1}, l_{i+1,2}, \ldots\ldots, l_{i+1,n_{i+1}}$
.	.
.	.
.	.
N(say)	$l_{N1}, \ldots\ldots\ldots, l_{Nn_N}$

The sampling procedure described above amounts to (i) selecting some of the N lines at random with equal probability 1/N and with replacement and (ii) if the i-th line happens to be selected, selecting the cluster of sentence-lengths ($l_{i-1,1} \ldots\ldots l_{i-1,n_{i-1}}$; $l_{i1} \ldots\ldots\ldots l_{in_i}$; $l_{i+1,1} \ldots\ldots l_{i+1,\,n_{i+1}}$) . Excepting for two lines, those at the two ends of the work, the probability of any given line being selected is clearly 3/N. So this procedure is giving equal probability of inclusion

to almost all the sentences in the work. But there is an element of clustering and the clusters are overlapping.

8.3.10. Incidentally, as in word-length studies, the probability sample with k clusters was spilt up into four independent and interpenetrating subsamples : clusters having order of selection 1, 2,.....k/4, gave subsample 1, those having order of selection k/4+1, ..., k/2 gave subsample 2, and so on.

8.3.11. Now define the i-th cluster as the cluster of sentences obtained through sample line i. Let x_i be the number of sentences in this cluster; then $x_i = n_{i-1} + n_i + n_{i+1}$. Let y_i be either the total length of all the sentences in the cluster or the number of sentences in the cluster having length equal to or less than or equal to some specified value l. We shall mostly use estimates (based on these probability samples) which are of the form $\frac{\sum y_i}{\sum x_i}$, where the summation is over the sample clusters. The large sample properties of such ratio estimates were discussed in Section 2.4 of Chapter 2 in connection with word-length studies. These ratios are consistent estimates of the corresponding ratios based on all the overlapping clusters in the population, which are, ignoring small differences, equal to the true values (i.e., average lengths or proportions or cumulative proportions of sentences) for the whole work. But the situation is not fully satisfactory if one is interested in using the other large sample properties of ratio estimates.

8.3.12. Consider the estimates given in Table 8.1 based on the entire probability samples from two selected works, viz.,'Gora' and 'Devayan'. The sample size for Gora is of the same order of magnitude as those

for most other works. Only 'Visavriksha' has an appreciably smaller size. 'Shakuntala', 'Sitar Vanavas' and 'Durgeshnandini' have larger sizes, because 200 randomly selected lines were used. The situation for these works and for the probability samples from 'Krishnakanter Will', 'Yogayog' and 'Kavi Shri Ramakrishna' seem to be represented by 'Devayan'. Secondly, Gora represents works with a high average of sentence-length and 'Devayan' those with a low average. These should serve to explain the choice of these two works for the calculations of C.V.

Table 8.1: Coefficients of variation of sample means of x and y, for some important ratio estimates of the form $\frac{\sum y_i}{\sum x_i}$, where the summation extends over all clusters of sentences included in the probability samples from 'Gora' or 'Devayan'.

variate (x or y)	averages		C.V.(%)		C.V. of sample mean (%) 1/2/	
	"Gora"	"Devayan"	"Gora"	"Devayan"	"Gora"	"Devayan"
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. total no. of sentences	3.258	4.062	17.4	24.0	2.21	2.66
2. no. of sentences with length=10	0.210	0.173	212.2	236.8	26.9	26.3
3. -do- with length ≤ 5	0.371	1.346	169.3	104.5	21.5	11.6
4. -do- -do- ≤ 15	2.177	3.568	54.4	48.0	6.91	5.34
5. -do- -do- ≤ 30	2.919	4.000	30.1	26.4	3.83	2.93
6. total length of all sentences	50.645	36.160	48.9	34.6	6.21	3.84

1/ No. of sentence-clusters : Gora -62, Devayan -83.

2/ C.V.'s of subsample means would be twice those of the combined means.

Table 8.2 Averages of sentence-length and estimates of parameters of lognormal distributions fitted to sentence-length distributions for twenty works in Bengali prose*.

sr. no.	author and work	type of sample	no.of sample sentences	average sentence-length in words: SS 1	SS 2	SS 3	SS 4	comb.	simple average of sub-sample averages	lognormal parameters: mean ($\log_{10}x$) ($\hat{\theta}$)	s.d. ($\log_{10}x$) ($\hat{\lambda}$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Vidyasagar										
1.	Shakuntala	prob.	220	15.65	14.96	14.67	14.79	15.01	15.02	1.101	0.240
2.	Sitar Vanavas	"	276	17.86	13.91	16.85	17.78	16.68	16.60	1.143	0.246
	Bamkimchandra										
3.	Durgeshnandini	"	297	11.68	11.03	11.40	9.60	10.89	10.93	0.923	0.304
4.	Visavriksha	"	164	8.26	8.92	10.13	8.70	8.98	9.00	0.850	0.277
5.	Krishnakanter Will	"	301	8.95	7.97	8.67	9.17	8.68	8.69	-	-
		syst.	390	8.07	10.19	10.92	8.13	9.28	9.32	-	-
		pooled	691					9.02		0.821	0.317
	Rabindranath										
6.	Gora	prob.	202	13.34	15.42	17.41	14.70	15.54	15.22	1.071	0.314
7.	Chaturanga (Pts.1-2)	complete count	675					12.43		0.992	0.294
8.	Sheser Kavita	prob.	192	11.15	10.24	11.66	11.08	11.02	11.03	0.928	0.330
9.	Yogayog	"	307	9.45	10.62	11.19	9.85	10.17	10.28		
		syst.	278	9.81	10.23	10.60	11.11	10.43	10.44		
		pooled	585					10.29		0.927	0.283

* The distributions are presented in Tables 8.3 to 8.5.

Table 8.2: (contd.)

sr. no.	author and work	type of sample	no. of sample sentences	average sentence-length in words					simple average of sub-sample averages	lognormal parameters	
				SS 1	SS 2	SS 3	SS 4	comb.		mean ($\log_{10}x$) ($\hat{\theta}$)	s.d. ($\log_{10}x$) ($\hat{\lambda}$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Pramatha Chaudhury										
10.	Char-Yari Katha	prob.	199	11.69	11.62	10.26	11.04	11.15	11.15	0.946	0.361
11.	Birbaler Halkhata	"	216	14.70	14.33	16.38	14.73	15.00	15.04	1.111	0.230
	Saratchandra										
12.	Pallisamaj	"	223	14.72	11.65	10.39	12.11	12.18	12.22	1.008	0.291
13.	Pather Dabi	"	203	13.92	13.38	11.70	11.93	12.69	12.73	0.996	0.341
	Bibhutibhusan										
14.	Pather Panchali	"	197	10.15	12.23	14.02	10.42	11.60	11.70	0.938	0.300
15.	Devayan	"	329	8.00	9.98	9.67	8.42	8.90	9.02	0.825	0.307
	Jajabar										
16.	Dristipat	"	209	8.20	10.20	10.31	8.96	9.43	9.42	0.851	0.318
17.	Janantik	"	242	9.36	9.29	9.68	7.52	8.90	8.96	0.851	0.334
	Muztaba Ali										
18.	Chacha Kahini	"	228	13.21	12.30	12.02	12.57	12.53	12.52	0.970	0.350
19.	Deshe Videshe	"	189	11.02	12.80	13.17	12.61	12.40	12.40	0.946	0.318
	Achintya Sengupta										
20.	Kavi Shri Ramakrishna	"	360	7.23	7.60	6.52	7.86	7.29	7.30		
		syst.	490	6.82	7.38	7.31	7.18	7.18	7.17		
		pooled	850					7.23		0.763	0.261

8.3.13. The number of sample clusters is well above 30 for the combined probability samples; also the C.V.'s in rows 1 and 6 under cols. (6) and (7) are well below 10%. So the estimated averages of sentence-length based on the combined probability samples may safely be assumed to possess the large sample properties (unbiasedness, normality etc.) of ratio estimates. The subsample averages may not possess these properties, however; the C.V.'s of sample $\bar{y}$ may be a little above 10%, and the number of clusters may be about 15 or 20. The estimates of p_1 (proportion of sentences of length 1) may be far from satisfactory even for the combined sample, since C.V.$(\bar{y})$ is of the over of 25%. This is because, in this case, the variate y is very often 0, and only occassionally 1 or more. Even the estimates of cumulative proportions $P_1 = p_1 + p_2 + \ldots\ldots + p_1$ do not seem to possess the large sample properties unless 1 is large so that P_1 is 40% or more, roughly speaking. The subsample estimates of P_1 would be satisfactory only for still higher values of 1, where P_1 exceeds 80%, roughly speaking. For consider the C.V. in row 4, col. (7). The cumulative proportion P_1 is nearly 85%, but C.V.$(\bar{y})$ is 5.34% for the combinedsample and hence above 10% for the subsamples.

8.3.14. Although the subsamples are rather too small in size, the subsample estimates of average sentence-length seem to practically unbiased. Compare the simple averages of the four subsample means given in col. (10) of Table 8.2 with the combined averages given in col. (9) of the same table. Out of the 19 rows for probability samples,

4 rows show the combined average larger, 13 others show the combined average smaller, while in the remaining two, the differences are practically zero. The differences are in most cases less than $\frac{1}{2}$ % ; the largest difference of about 2% is seen for Gora. The average difference is about 0.005. From the discussion in Chapter 2, Section 2.5, on the significance of these differences, it follows that the bias of the estimates is on the whole negligible, especially for the combined sample.

8.4.1. Sentence-Sampling : As in word-length studies, again systematic samples were also taken from three selected works, and some experiments conducted, to examine the possibility of using 'systematic' samples of lines for giving valid samples of sentences. The definition of a 'systematic' sample may best be given through examples. We took, for instance, the 4th line from bottom of every alternate page, and sample the sentence(s), if any, terminating on these lines. No extension was done to preceding or following lines having termination marks. Four such rules were followed for sampling from each of the three works, and this gave four independent and interpenetrating subsamples. The matter will be reported in detail in the next chapter, but the 'systematic' samples, found to closely resemble probability samples, will be used in the present chapter.

8.4.2. Some observations on sentence sampling may perhaps be made here. Complete enumeration of entire works is far less time-consuming than in the case of word-length studies, provided the difficulties

of defining sentences are not serious. About half, actually Parts 1 and 2, of Tagore's short novel, "Chaturanga", was counted completely in only a few hours, and even then one got only 675 sentences. Fortunately, even the small-sized probability samples taken here suffice for drawing many important conclusions. If larger samples are needed, one should not waste one's time by following the method of probability sampling adopted here. One should at least extend the cluster to a greater extent, e.g., when the i-th line is selected, one may take say 5 lines (having terminations) on either side to form the sentence-cluster. [Vide para 8.3.8 above. A similar observation on word sampling was made in Section 2.3, Chapter 2.] Systematic sampling may also be used, if not complete enumeration.

8.4.3. If one is interested only in the average of sentence-length, one may use the random samples of lines[1] of Chapter 2 more directly. For the i-th sample line selected, let y_i be the number of words on the line and x_i the number of termination marks on it. Here x_i may be zero, that is, lines not having any termination mark are not being left out. One can then use the ratio estimate $\frac{\sum y_i}{\sum x_i}$, the summation extending over all lines, provided the number of sample lines is greater than 30 and also sufficiently large to make the C.V.'s of sample means $\bar{x}$ and $\bar{y}$ less than 10%.

8.4.4. To digress a little at this point, the above method may be extended for studying the relative frequencies of different types of punctuation marks per 100 or 1000 running words. Let y_i remain as

1/ Or clusters of a number of consecutive lines.

before, but x_i be the number of commas, say, in the i-th sample line. Then one may use $\frac{\sum x_i}{\sum y_i}$ where the summation is over all the sample lines or line-clusters. Such data are presented by Dewey (1923) and Miller (1951, Chapter 6) for English. Miller makes some interesting observations on the same pointing out their uses for studies on style and also for studying historical trends in punctuation.

8.5.1. Sentence-length in Bengali prose - a broad survey : Tables 8.3 to 8.5 present many sentence-length distributions. Table 8.3 covers the sixteen works in Bengali prose which were subjected to probability sampling alone, and Table 8.4 three other works in Bengali prose subjected to both probability and systematic sampling. In both tables, the distributions are shown separately by subsamples as well as for the combined sample. Table 8.5 gives the sentence-length distribution based on a complete count of Parts 1 and 2 of Tagore's novel, "Chaturanga". It gives besides the distributions for three short essays in Bengali, one extract from Tagore's novel, "Sheser Kavita", and two extracts from Jane Austen's (English) novel, "Pride and Prejudice". All distributions in Table 8.5 are based on complete counts and some of them are needed only in Chapter 9. Table 8.2 presents the subsamplewise and combined averages of sentence-length for the nineteen works covered in Tables 8.3-8.4 and for "Chaturanga" covered in Table 8.5. But Table 8.5 includes the averages for all the distributions shown in it. Table 8.2 also presents the estimates of parameters of log-normal distributions fitted, fairly successfully, to the sentence-length distributions (vide Section 8.7).

Table 8.3: Distributions of sentence-length in sixteen works in Bengali prose, based on probability samples of sentences[1/]

length (no. of words)	no. of sentences by works and by sub-samples: Shakuntala ss 1	ss 2	ss 3	ss 4	comb.	Sitar Vanavas ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1										
2							1			1
3			2	1	3		1			1
4		1	1		2	2	1			3
5	1	4	1	3	9		2			2
6			1	2	3	2	4		3	9
7	1	2	1	5	9	3	3	2	3	11
8	6	3	3	1	13	2	6	4	1	13
9	7	4	6	7	24	5	3	7	6	21
10	3	4	5	1	13	1	6	1	5	13
11	5	5	1	2	13	3	4	5	7	19
12	3	4	5	3	15	8	4	4	7	23
13	2	2	3	3	10	4	3	1	6	14
14	3	3	1	4	11	1	2	1	4	8
15	3	4	4	3	14	4	5	1	4	14
16	4	1	1	3	9	3	5	1	4	13
17			1	3	4	3		4	4	11
18	2	1	2	2	7	2	3	5	3	13
19	1	5	2	2	10	2	3	4	1	10
20	2		3	1	6	3	1	2	1	7
21			3	1	4	5		3	3	11
22	1	3	2	3	9	1	2	1	3	7
23	1	1	2		4	1	2		5	8
24	3	1	1		5	3	1		2	6
25		1	1		2	2	2	2	1	7
26	1	1		2	4					
27	1	1		1	3					
28				1	1			1	1	2
29		1			1	1		1		2
30	1		1	2	4	3				3
31				1	1	1			1	2
33	1				1	1				1
34									2	2
35						1		1		2
36				1	1		1		1	2
37		1			1			1	1	2
38	1				1				2	2
39						1	1			2
40									2	2
43								1		1
49							1	1	1	3
53		1	1		2	1				1
58	1	1			2	1			1	1
68						1				1
total	54	54	54	58	220	70	67	54	85	276

1/ No. of lines selected at random : Shakuntala-200, Sitar Vanavas-200
No. of clusters : Shakuntala-73, Sitar Vanavasa-90.

Table 8.3: (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Durgeshnandini					Visavriksha				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
1		1	3		4					
2	3		1	2	6	1	1		3	5
3	3	3	4	4	14	3	2	3	4	12
4	5	5	5	8	23	5	4	2	3	14
5	8	4	5	7	24	5	5	2	4	16
6	2	5	1	11	19	2	2	4	5	13
7	8	8	1	9	26	5	4	3	4	16
8	3	6	5	4	18	4	5	5	3	17
9	6	7	8	3	24	4	3	3	3	13
10	3	3	8	3	17	1	3	3	4	11
11	1	4	5	6	16	4	3	2	1	10
12	1	5	4	2	12	3				3
13	2	5	2	4	13		3	1		4
14	5	2	1	5	13		1	3	4	8
15	3	2	2		7	1		1		2
16	2	2	1	4	9	1		2		3
17	1	1	3	2	7	1		2	1	4
18		1			1	1	2			3
19	6	4			10					
20	1	1	1		3	1				1
21	4		3		7			1	1	2
22		1	1		2				1	1
23				3	3		1	1	1	3
24	2	2			4			1		1
25	1				1				1	1
26				1	1		1			1
27	1		1		2					
28				1	1					
29			1		1					
30			2		2					
31	1	1			2					
34			1	1	2					
36			1		1					
40	1				1					
48		1			1					
total	73	74	70	80	297	42	40	39	43	164

1/ No. of lines selected at random : Durgeshnandini-200, Visavriksha -100.
No. of clusters : Durgeshnandini -94, Visavriksha- 49.

Table 8.3: (Contd.)

length (no. of words)	no. of sentences by workband by sub-samples									
	Gora					Sheser Kavita				
	ss 1	ss 2	ss 3	ss 4	comb.	ss.1	ss 2	ss 3	ss 4	comb.
(1)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)
1						2				2
2	1				1	3	4		1	8
3	2	1	2	1	6	3	3		2	8
4		1	4	2	7	4	7	4	3	18
5	2	3	2	2	9	1	3	3	4	11
6		2	4	2	8	1	4	2	4	11
7	2	2	4	1	9	5		3	3	11
8	4	3	5		12	2	5	3	3	13
9	1	6	6	3	16	2	1	5	5	13
10	4	3	2	4	13	2	4	5	3	14
11	2	3	2	4	11	3	1	3	4	11
12	5	4	2	3	14	5	2			7
13	2	2	1	3	8	2	3	3	2	10
14		3	6	4	13	1	3	2	1	7
15	1	2	2	3	8	1	2	3	5	11
16	1	3	1	1	6	1	3	1	1	6
17	3		1	3	7	1	1	1		3
18	2	1	3		6			3	2	5
19		2	2	1	5	1		1		2
20		1	2		3	2				2
21	1	1		2	4			1	1	2
22	2				2	2	1	2		5
23		1	2		3					
24		1			1			1	1	2
25	1		2		3		1	1		2
26				2	2		1		1	2
27			1		1					
28		1			1					
29				1	1					
30			1		1					
31			3		3					
32	1	1			2		1			1
33		1			1		1		2	3
34		1			1					
36		1			1					
37		1			1					
39			1		1					
40			1		1					
42			1		1					
44			1		1					
46	1		1		2					
47						1				1
48			1		1					
49						1				1
50			1		1					
51				1	1					
60				1	1					
75		1			1					
93			1		1					
total	38	52	68	44	202	46	51	47	48	192

1/ No. of lines selected at random : Gora -100, Sheser Kavita - 100.
No. of clusters : Gora - 62, Sheser Kavita - 60.

Table 8.3 : (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Birbaler Halkhata					Char-Yari Katha				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(32)	(33)	(34)	(35)	(36)	(37)	(38)	(39)	(40)	(41)
1						3	1	3	2	9
2	1				1	1	1		1	3
3		1		1	2	2	2	3	4	11
4	1			1	2	1	3	1	4	9
5	1	2		2	5	3	5	3	4	15
6	1	2		6	9	3		1	4	8
7	1	3	2	2	8	6	3	5	3	17
8	9	5	3	2	19	5	6	5	2	18
9	6	3	1	3	13	3	2	2	8	15
10	3	3	8	6	20	2	1	7	1	11
11	2	2	1	1	6	3	4	2	1	10
12	1	5	2	2	10	2	4	1	3	10
13	1	4	1	1	7	1	1	4	2	8
14	2		5	6	13			2	4	6
15	3	3	1	6	13	1	2			3
16	2	3	1	1	7	1		1	1	3
17	5	2	2	1	10	4	1	1	2	8
18	1		2	2	5		1		1	2
19	3	2	3	2	10		2	1	2	5
20		1	8	2	11	1	2	1	3	7
21	1	1	1		3			2		2
22	2	5	3	1	11	2	1			3
23		2	1	2	5					
24	1		1	2	4			1		1
25	2	1	1		4	1	2			3
26	1	1			2	1			1	2
27				1	1		1		1	2
28		3	1	2	6		2			2
29				1	1					
30	1		1	1	3	1		1		2
31				1	1	1				1
36									1	1
39			1		1				1	1
41				1	1					
42	1				1					
50	1				1					
51						1				1
total	53	54	50	59	216	49	47	47	56	199

1/ No.of lines selected at random: Birbaler Halkhata -100,
Char-Yari Katha - 100.
No. of clusters : Birbaler Halkhata - 66, Char-Yari Katha -62.

Table 8.3: (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Pallisamaj					Pather Dabi				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(42)	(43)	(44)	(45)	(46)	(47)	(48)	(49)	(50)	(51)
1			1	1	2		1	2	1	4
2		2	1	1	4			3	3	6
3		2	3	1	6	2	1	2	3	8
4	1	3	5	3	12	2	2	2	2	8
5	4	6	5	3	18	2	2	7	1	12
6	5	1	3	2	11	3	1	5	5	14
7	1	4	5	3	13	5	1	3	6	15
8	1	4	7	6	18	2	4	1	3	10
9	6	5	1	3	15	3	4	1	6	14
10	4	4	2	3	13		5		1	6
11	3	2	7	4	16	4	3	5	4	16
12	2	2	1	1	6	8		2	3	13
13	3	2		3	8	1	2	1	2	6
14	2	2	1	4	9	1	1	2	1	5
15	3	1	2	2	8	2	5	1	1	9
16		4	4	1	9	1	1		1	3
17	3	2	1	5	11		1	1	1	3
18	3			1	4	1	1	2		4
19	1	3	1	2	7	3	2	2		7
20	1	1	2		4		3		1	4
21	2	1	2		5	1	3		3	7
22		1		2	3	1		1	1	3
23	1	3		1	5			1	1	2
24				1	1			1	1	2
25						1	1		1	3
26	2	1	2		5		1		1	2
27				1	1			2		2
28						2			1	3
29	1				1				2	2
30	1		1		2	1				1
31							1			1
33		1			1					
34	2				2		1	1	1	3
36	1				1	1				1
38								1		1
39	1			1	2	1		1		2
45						1				1
total	54	57	57	55	223	49	47	50	57	203

1/ No. of lines selected at random : Pallisamaj - 100, Pather Dabi - 100.
No. of clusters : Pallisamaj -68, Pather Dabi - 62.

Table 8.3 : (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Pather Panchali					Devayan				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(52)	(53)	(54)	(55)	(56)	(57)	(58)	(59)	(60)	(61)
1	1		1	2	4	1	•	1	2	4
2		1			1	6	3	1	6	16
3		1	1	3	5	11	5	5	9	30
4	5	2	2	3	12	6	10	10	6	32
5	5	5	2	2	14	12	4	6	5	27
6	6	9	6	3	24	11	6	9	9	35
7	4	1	6	6	17	6	6	4	7	23
8	4	3	1	4	12	8	3	9	10	30
9	4		2	8	14	4	2	1	9	16
10	4	5	3	1	13	4	4	3	3	14
11	3			4	7	5	2	6	5	18
12	2	1	6	1	10	7	5	2	4	18
13	3	3	1		7	2	1	5	3	11
14	2	2	1	2	7	1	1	1	2	5
15	1	2	1	4	8	3	2	3	2	10
16	2	2	1	2	7	1	2	3	3	9
17			1	3	4	2	1	2	3	8
18	1	1		2	4	1	1	1		3
19	1	2			3		1		1	2
20	1	1			2	1		1		2
21	1		1		2	1		1	1	3
22				1	1					
23						1	1			2
24		2	1		3			1		1
25			1		1		1	1		2
26	1	2	1		4	1			2	3
27			1		1					
28										
29		1		1	2					
30			1		1					
31			1	1	2					
32			1		1					
33								1		1
34			1		1					
35	1				1					
36							1	1		2
39							1			1
50							1			1
56		1			1					
78			1		1					
total	52	47	45	53	197	95	64	78	92	329

1/ No. of lines selected at random : Pather Panchali -100, Devayan-100.
No. of clusters : Pather Panchali -58, Devayan - 83.

Table 8.3 : (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Dristipat					Janantik				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(62)	(63)	(64)	(65)	(66)	(67)	(68)	(69)	(70)	(71)
1	1	1	1	1	4		2	1	2	5
2	2	2	2	2	8	3	4	2	4	13
3	3	2	6	2	13	2	2	5	7	16
4	6	2	4	7	19	3	1	2	8	14
5	9	5	4	2	20	5	5	2	7	19
6	2	2	3	4	11	7	9	6	6	28
7	5	4	5	5	19	10	2	9	6	27
8	5	3	6	3	17	4	7	6	3	20
9	3	6	1	5	15	7	4	4	7	22
10	1	8	3	3	15	1	3	2	4	10
11	2	3	4	4	13	1	2	5	1	9
12	2	3	2	7	14	3	1	2	1	7
13	2	2	1	2	7	2	3	1	2	8
14				1	1	1	1	3	3	8
15	3				3	2		1	2	5
16	1	1	2	2	6		1	2	1	4
17		2	1		3	1	3		1	5
18	1		2	1	4	3	2	1	3	9
19	1	2			3					
20	1				1	1	1	1		3
21	1		1		2			1		1
23							1	1		2
24			2	1	3	2				2
25		1			1		1	1		2
26				1	1					
27							1			1
28			1		1					
29			1		1					
30			1		1	1				1
33		1			1					
36			1		1					
37		1			1					
42								1		1
total	51	51	54	53	209	59	56	59	68	242

1/ No. of lines selected at random : Dristipat -100, Janantik -100.
No. of clusters : Dristipat -60, Janantik -68.

Table 8.3 : (Contd.)

length (no.of words)	no. of sentences by works and by sub-samples									
	Chacha-Kahini					Deshe-Videshe				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(72)	(73)	(74)	(75)	(76)	(77)	(78)	(79)	(80)	(81)
1	1	1		2	4		1			1
2	2	1	4	1	8	2		1		3
3	3	4	1	3	11	2		5	4	11
4	4	5	5	3	17	5	3	1	7	16
5	4	3	3	4	14	4	4	6	2	16
6		5	4	5	14	2	4	2	3	11
7	2	4	4	3	13	1	3	3	2	9
8	1	6	3	2	12	4	3	2	3	12
9	5	1	4	5	15	1	4	2	3	10
10	1	5	3	1	10	5	1	2	2	10
11	5	2	2	2	11	2	3	1	1	7
12		2	2	2	6	1	2	4	3	10
13	4	1	2	4	11	1	1		2	4
14	3		2	4	9	4	1	2	3	10
15	2	1	1	1	5	1	1	1	6	9
16	2	3	2	2	9	2	1	1	1	5
17	3		1		4	3	1	1	2	7
18	1		1		2	1	3	2		6
19	1	2	2	1	6		2	2	1	5
20	2				2	2	1	1		4
21	3	1	3		7	1		3	1	5
22	1	2		1	4		2			2
23			2	3	5	1			1	2
24	3		2	2	7					
25	1				1				1	1
26	1	1	1		3			1	1	2
27	1	2	1		4	1		1		2
28	1	1	1	1	4	1			1	2
29							1			1
31		1			1					
33				1	1					
34		1			1					
36								1		1
37	1				1					
38		1		1	2					
39			1	1	2					
40				1	1		1			1
50		1			1			2		2
58							1			1
93									1	1
total	58	57	57	56	228	47	44	47	51	189

1/ No. of lines selected at random : Chacha-Kahini - 100, Deshe-Videshe - 100.
No. of clusters : Chacha-Kahini -68, Deshe-Videshe -59.

Table 8.4: Distributions of sentence-length in three works in Bengali prose, subjected to both probability sampling and systematic sampling

(a) Krishnakanter Will.

length (no. of words)	no. of sentences by type of sample and by subsamples									
	probability sample					systematic sample				
	SS 1	SS 2	SS 3	SS 4	comb.	SS 1	SS 2	SS 3	SS 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	1	2	1	2	6	3		1		4
2	3	5	3	3	14	3	1	8	4	16
3	8	9	7	8	32	8	5	7	9	29
4	7	10	5	2	24	14	9	9	11	43
5	4	9	6	9	28	9	7	10	13	39
6	6	7	7	11	31	11	14	7	11	43
7	6	6	10	5	27	7	4	7	9	27
8	8	4	5	5	22	5	3	4	10	22
9	4	4	1	4	13	6	13	4	4	27
10	7	2	8	3	20	5	4	7	11	27
11	3	3	2	2	10	6	7	2	8	23
12	2	3	2	3	10	2	6	4	6	18
13	1	1	2	5	9	3	2	4		9
14	3	1	4	3	11	1	1	3	7	12
15	2	3	2	3	10	3	2	1	1	7
16		3	0		3	1	1	1	2	5
17	1	1	1	1	4		1	1	3	5
18	2	1		3	6		1	2		3
19	2	1	2		5		1	1		2
20			2	1	3			2	1	3
21		1			1	1	1	2		4
22		2		1	3			1		1
23			1	1	2		2			2
24	1				1		2			2
25								1		1
26		1			1	1				1
27	1				1			2	1	3
28						1				1
29			1	0	1			1		1
30						1				1
31								1		1
33				2	2					
34							1			1
38						1				1
39								1		1
40								2		2
41	1				1					
42							1			1
43							1			1
51								1		1
72										
total	73	79	72	77	301	92	90	97	111	390

Table 8.4: (contd.)

(b) Yogayog

length (no. of words)	no. of sentences by type of sample and by subsamples									
	probability sample					systematic sample				
	SS 1	SS 2	SS 3	SS 4	comb.	SS 1	SS 2	SS 3	SS 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1			2		2			1	1	2
2	1	1	3	1	6	4	2	1	2	9
3	6	1	3	4	14	4	9	1	2	16
4	5	4	5	4	18	6	5	2	6	19
5	14	7	3	3	27	4	3	5	3	15
6	4	9	3	6	22	4	5	9	4	22
7	6	8	10	5	29	5	8	5	3	21
8	6	8	3	8	25	10	5	3	8	26
9	5	5	8	9	27	2	3	3	7	15
10	3	3	4	5	15	6	3	2	5	16
11	5	3	3	5	16	5	6	4	5	20
12	6	5	3	3	17	2	7	4	3	16
13	3	6	6	5	20	2	5	2	3	12
14	5	3	2		10	2	1	1	3	7
15	2	3	3	2	10	3	2	3	4	12
16	2	3	4	3	12	2	4	4	1	11
17	1	2		3	6		4	1	2	7
18	2		2	1	5	3	1	1	1	6
19	1	4			5		2	1	1	4
20	2	2	1	1	6				3	3
21	1			2	3	1			1	2
22	1	1			2		1		1	2
23	1			1	2	1		1		2
24			2		2		1			1
25										
26			2		2	1		1		2
27									1	1
28		1			1		1	1	1	3
29						1				1
30									1	1
31			1		1					
32									1	1
33						1		1		2
37							1			1
39		1			1					
50			1		1					
total	82	80	74	71	307	69	79	57	73	278

Table 8.4: (contd.)

(o) Kavi Shri Ramakrishna

length (no. of words)	no. of sentences by type of sample and by subsamples									
	probability sample					systematic sample				
	SS 1	SS 2	SS 3	SS 4	comb.	SS 1	SS 2	SS 3	SS 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1		1			1	1	1	1		3
2	5	6	6	3	20	3	4		7	14
3	11	9	10	5	35	12	15	13	8	48
4	16	9	14	8	47	14	18	16	9	57
5	8	14	14	9	45	22	15	23	16	76
6	5	7	10	15	37	11	18	16	18	63
7	7	12	13	7	39	11	17	9	17	54
8	7	10	4	9	30	10	9	5	9	33
9	7	4	6	6	23	8	8	7	9	32
10	5	3	3	6	17	6	8	5	11	30
11	4	6	4	3	17	8	2	4	5	19
12	2	1	4	4	11	2	3	8	4	17
13	2	2		1	5	2	6		1	9
14	2			3	5	1	3	5	2	11
15	3	4	3	3	13		3		1	4
16	1	1	1	1	4	2		3	1	6
17				1	1		1	2		3
18	1		1		2	2				2
19	1				1		1			1
20							2	1		3
21	1				1				1	1
22		2			2				1	1
23		1			1					
24		1			1		1			1
25		1			1					
27								1		1
29							1			1
31				1	1					
total	88	94	93	85	360	115	136	119	120	490

Table 8.5: Sentence-length distributions based on complete counts of selected works or parts of works in Bengali and English prose.

length (no.of words)	no. of sentences by works								
	Chaturanga			Sheser Kavita: extract	Samya	Bamkim-chandra	Vishwa-vidya-lay	Pride and Prejudice	
	Part 1	Part 2	comb.					extract 1	extract 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1		2	2		2				
2	6	3	9	2	2			1	
3	9	13	22	5	6	2			
4	21	23	44	3	9	1	4		
5	24	20	44	9	17	4	3	1	2
6	24	23	47	7	18	2	4	4	2
7	22	20	42	9	15	4	6	4	5
8	26	19	45	11	17	7	8	1	4
9	19	20	39	10	7	9	7	3	1
10	32	17	49	6	8	7	7	4	5
11	27	12	39	8	9	12	9	3	3
12	13	12	25	14	8	8	7	1	10
13	14	7	21	5	5	4	8		7
14	16	11	27	5	3	12	5	5	3
15	22	9	31	6	6	4	3	1	5
16	11	15	26	7	6	7	5	3	3
17	8	7	15	7	6	5	4	3	7
18	14	9	23	4	3	3	6	1	3
19	8	11	19	5	10	2	3	2	6
20	13	6	19	7	5	5	2	3	6
21	8	4	12	3	2	7	4	5	6
22	4	4	8	3		5	3		5
23	8	2	10	1	4	3	7	3	1
24	3	3	6	2		1	1	1	5
25	3	1	4	2	1	5	2		9
26	3	2	5	1	2	2	3	2	4
27	5	4	9		1	2	2	1	4
28	4	3	7	2	1	1	1	1	3
29	1		1	3	1	1		5	4
30	3	1	4	2	2	1	1	6	3
31	3	1	4	1	1	2	2	3	2
32		3	3	1	1	1	1	3	3
33	3	1	4	1				2	2
34						1	1	1	6
35		1	1			1	2	4	6
36-	3[a]	6[b]	9	3[c]	1[d]	8[e]	4[f]	29[g]	54[h]
total	380	295	675	155	179	139	125	106	189
average	12.51	12.32	12.43	13.81	11.34	17.41	16.16	28.25	29.45

a/ 36,37,51 b/ 38,39,41,45,63,100 c/ 36,37,55 d/ 40
e/ 38,38,41,44,55,56,68,75 f/ 44, 46, 47, 61
g/ 37,37,37,38,38,39,39,39,40,40,40,41,42,42,43,44,46,47,48,52,53,55,56, 58,60,63,64,105,114
h/ 36,36,37,37,38,39,39,39,41,41,42,43,43,43,44,44,45,45,46,46,47,48,48, 49,49,49,51,51,51,51,52,52,53,53,54,55,55,56,56,58,59,60,60,60,64,65, 66,71,73,73,74,78,81,83.

8.5.2. One may first look at Table 8.2 presenting the subsample-wise and combined averages of sentence-length in twenty works in Bengali prose. The average length is smallest, about 7.25, for "Kavi Shri Ramakrishna", which belongs to religious literature. This work has a very unusual style. One striking feature is the unusual vocabulary. But that is not all : very short sentences is the other important characteristic.

8.5.3. The average is almost 9 for a number of works, viz., "Visavriksha","Krishnakanter Will", "Devayan", "Dristipat" and "Janantik"; but 11 or 12 seems to be the modal value of the average for Bengali prose fiction.

8.5.4. The higher values among the averages may now be noticed. "Gora" is loaded with elevated discussions and argumentation on socio-religious topics. This explains the comparatively high average, about 15.5. A similar high figure (15.0) is shown by "Birbaler Halkhata", the collection of essays. "Char-Yari Katha" is a fiction by the same author in the same "Chalita" (colloquial) style, and shows an average of only about 11. The two essays by Tagore, viz., "Bamkimchandra" and "Vishwavidyalay", also show similar high averages (vide Table 8.5). Apparently essays tend to have averages of the order of 15, or at least, averages appreciably larger than those for ordinary fiction. This must be partly due to the presence of conversational matter in fiction, which tends to introduce shorter sentences. But certain works of fiction, like "Gora", resemble essays in having :

long sentences, while particular essays, like "Samya" by Bamkimchandra (vide Table 8.5), may approach fiction by having sentences as short as in ordinary fiction[1].

8.5.5. The high averages for "Shakuntala" and "Sitar Vanavas" are somewhat deceptive. As already stated, there are many semicolons in these two works, and quite a few seem to be inappropriately used, in the place of full-stops. /The Bengali equivalent of the full-stop is the vertical stroke./ If one ventures to revise the punctuation on the sample lines, one would get much smaller averages : For both the works, the averages could come down to 12, roughly speaking, if semicolons were avoided as far as possible. It may be mentioned in this connection that Vidyasagar was one of the makers of modern Bengali prose, and was therefore, one of the first few to use the western system of punctuation in Bengali writing. /Bengali had very little of punctuation originally, beyond the equivalent of the full-stop./ Some incorrect use of the semicolon should not be surprising in such circumstances, even when the author is as fastidious as Vidyasagar was. A look at the texts of "Shakuntala" and "Sitar Vanavas" would convince anyone that Vidyasagar used the comma also far too frequently.

1/ The estimates of θ i.e. of mean $(\log_{10} x)$ lead to more or less the same type of conclusions as given here. This is as expected, for mean $(\log_{10} x)$ is the logarithm of the median or geometric mean of the hypothetical continuous variate underlying the discrete sentence length values, while the observed mean is only slightly larger than the mean of this underlying variate.

8.5.6. The range of variation in the average sentence length, from about 7 to over 15, is much larger in relative terms than the range of variation in average word-length. This corroborates the statements by Williams (1940) and others that sentence-length is likely to be a more sensitive indicator of style than word-length.

8.5.7. It may be of interest to quote some figures for sentence-length in other languages. Flesch (1946, Chap. 4, especially, p.38) says that in English prose, an average length of 17 words seems to be the standard; literary English (fairly difficult) often uses an average of 21, and scientific English (very difficult) about 30; "very easy" prose with many dialogues is often written in 8-word sentences, on the average. Since 100 English words generally correspond to 75 or 80 Bengali words (Appendix 6), literary English seems to use about equally long sentences on the average as Bengali essays. And English essays seem to use somewhat longer sentences, on the average, than essays in Bengali. In the study by Yule (1938) based on serious essays by English writers the averages ranged from 22 for Macaulay's "Essays" to about 60 for "Economic Writings" by Sir William Petty; sentences with 200 or 300 words were occasionally found. Again, for English essays by Chesterton, Wells and Shaw, the averages were found to be between 25 and 30 (Williams, 1940). For some Latin works by Kempis and Gerson (Yule, 1938) the averages ranged from 15 to 23; for the Kannada works considered by Subba Rao (1960), the averages were very small, about 7 or 8. But it would be unsafe to generalise about these two languages from such meagre statistical evidence.

8.5.8. It is possible to study other characteristics of the distributions, like median or quartiles or deciles, as done by Yule (1938) or standard deviation and coefficient of variation. These characteristics could be obtained by subsamples. No such attempt has been made here because, since the discrete sentence-length distributions can be regarded as grouped versions of continuous lognormal distributions, to a fair degree of approximation, the two parameters of the underlying lognormal distribution, viz., E(log x) and var (log x) together summarise almost the whole information contained in the distribution [vide Williams (1940)]. The variation in mean (log x) has been seen already, though indirectly, through the averages $\bar{x}$. The variation in s.d.(log x) may now be examined.

8.5.9. As stated by Williams (ibid), the s.d.(log x) figures throw new light on the material. They indicate the C.V. of the underlying continuous variate and therefore, to a fair approximation, that of the discrete sentence-length values. Not only do the works vary in respect of the average, but they also vary considerably in respect of the C.V. The following points emerge from the figures for s.d.($\log_{10}$x), presented in Table 8.2.[1/]

8.5.10. The C.V. is low, of the order of 60%, for three works with high averages, viz., "Shakuntala", "Sitar Vanavas" and "Birbaler Halkhata" but for "Gora" the C.V. is quite moderate, about 80%. The

1/ If the estimate of s.d.($\log_{10}$x) be denoted by $\hat{\lambda}$, the C.V. is approximately equal to $\sqrt{e^{\hat{\lambda}^2(\log_e 10)^2} - 1}$

C.V. is low for "Kavi Shri Ramakrishna", the work with the lowest average. But the C.V. is high, about 95%, for "Char-Yari Katha", "Pather Dabi" and "Chacha Kahini" all having medium values of the average length. On the whole, the two aspects, viz., average and C.V., do not show any appreciable correlation.

8.5.11. Both these aspects can be seen together in Fig.8.1 which shows the values of $\hat{\theta}$ and $\hat{\lambda}$ i.e., mean $(\log_{10} x)$ and s.d.$(\log_{10} x)$ for different works, each work being represented by a point. Each point is liable to some errors due to sampling.[1/] The classification shown in Table 8.6 has, however, been attempted, to get a rough summary of the situation. It is possible that some works have been wrongly classified, depending on estimates based on rather small samples.

8.5.12. The C.V. of sentence-length is a subtler style-measure than the simple average; but unfortunately it seems to be more sensitive to the defects in punctuation, like using semicolons incorrectly where fullstops should be given, especially those giving rise to long sentences.

1/ Ignoring the small deviations from random sampling, but taking into account the method of estimation adopted in the present case, the standard errors of the estimates should be nearly 0.025 for $\hat{\theta}$ and 0.02 for $\hat{\lambda}$ for works from which about 200 sentences were sampled.

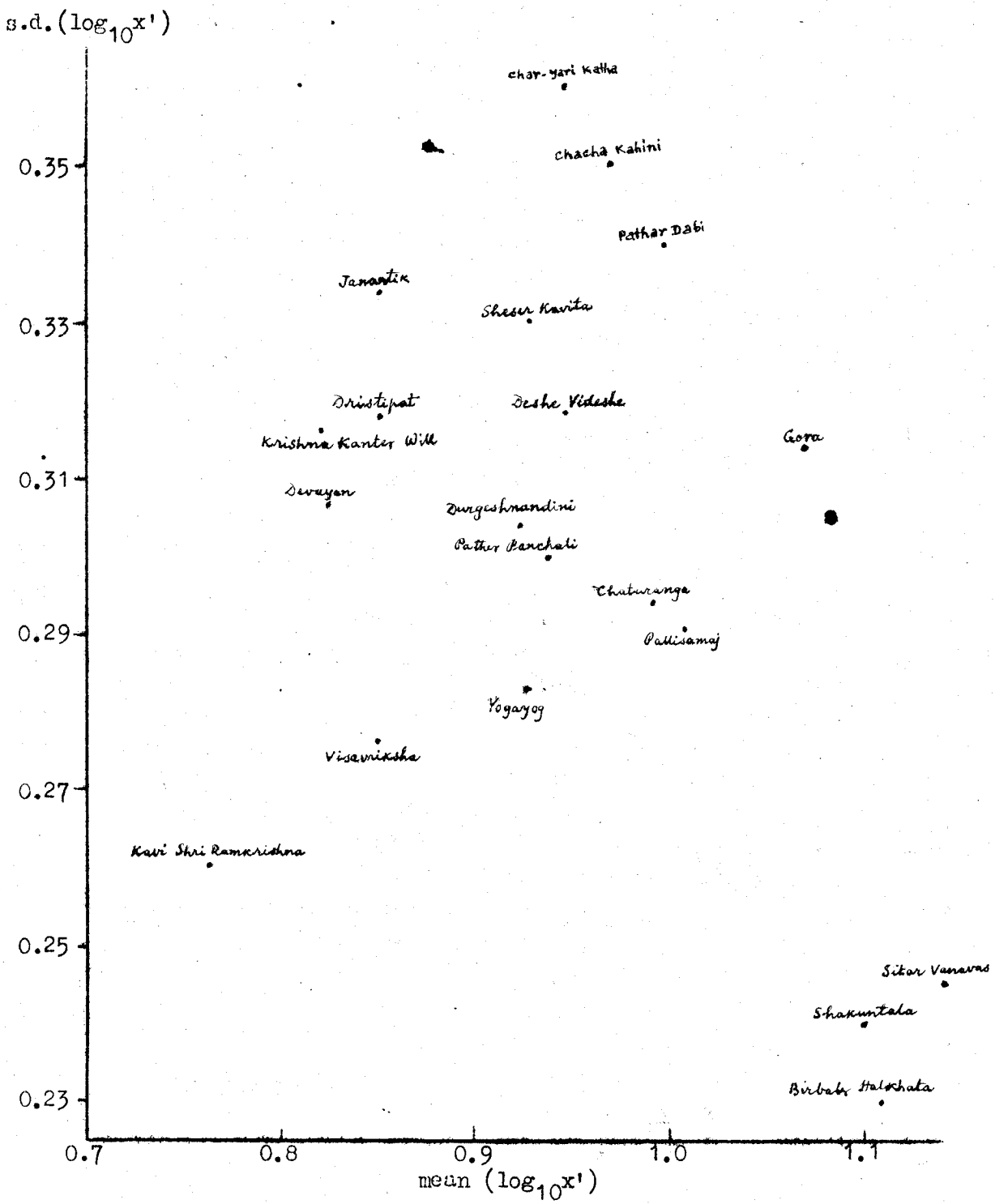


Fig. 8.1: Estimates of parameters of lognormal curves fitted to distributions of sentence-length in terms of words (x) for twenty works in Bengali prose, assuming that observed values of x, e.g., 1,2,3,...., represent intervals 0-1, 1-2, 2-3, ..., of the underlying lognormal variate (x').

Table 8.6: A classification of the twenty works covered in Table 8.2 according to average and coefficient of variation (C.V.) of sentence-length.

average sentence-length	coefficient of variation of sentence-length*			
	high	medium	low	very low
(1)	(2)	(3)	(4)	(5)
high (above 13)		Gora		Shakuntala, Sitar Vanavas, Birbaler Halkhata
medium (10-13)	Char-Yari Katha, Pather Dabi, Chacha Kahini	Durgeshnandini, Sheser Kavita, Deshe Videshe	Chaturanga (Pts. 1-2), Yogayog, Pallisamaj, Pather Panchali	
low (8-10)		Krishnakanter Will, Devayan, Dristipat, Janantik	Visavriksha	
very low (<8)				Kavi Shri Ramakrishna

* In terms of s.d.($\log_{10}x$) of the fitted lognormal distribution, the four categories represent intevals 0.34-0.37, 0.30-0.34, 0.27-0.30 and 0.23-0.27.

8.5.13. It is interesting to note that Bamkimchandra seems to have used medium or rather short sentences, on an average, contrary to popular impressions coloured by the elevated style. Also, if Bamkimchandra's figures are remembered along with those for Vidyasagar, there seems to have been little or no time-trend in the average sentence-length in Bengali prose. A declining trend, it may be recalled, was found in the average of word-length (vide Chapter 3).

8.6.1. Within author differences : The coverage of works is admittedly inadequate for one to draw inferences about between author differences; but the question whether and how far sentence-length can serve as a statistical index of individual style can be decided by examining within author differences.

8.6.2. The answer to the question is largely in the negative. Considerable and significant within author differences have been observed even in this study, where only a few works of each author have been covered. The differences might be attributed to changes in the author's style with increasing age and also with the topic or subject-matter or elevation of the work; but then the statement that sentence-length can indicate an author's style would have very little content and usefulness.

8.6.3. Compare, for example, the subsample-wise averages given in Table 8.2 for "Gora" and "Sheser Kavita". The average for "Gora" is higher, appreciably, in all the four subsamples. The difference between the overall averages, 15.54 for "Gora" and 11.02 for "Sheser Kavita" must therefore be regarded as significant.[1/] The average (12.43) for "Chaturanga", Parts 1 and 2, counted complete, seems to be significantly higher than the average (11.02) for "Sheser Kavita" and significantly lower than the average (15.54) for "Gora", "Yogayog" (10.29) seems to have an even lower average than "Sheser Kavita", but the difference may not be significant.

1/ The sign test is used here. Since the alternatives are one-sided, the level of significance is $6\frac{1}{4}\%$.

8.6.4. The technique of fractile graphical analysis (FGA) was employed for testing the significance of within author differences. Table 8.7 and Fig.8.2 are presented for illustrating the technique. These show the sentence-length distributions for two works by Tagore in the form of decile group averages of sentence-length. For each work, the four subsamples were grouped to form two halfsamples, for the sake of convenience. The fractile graphs for "Sheser Kavita" are uniformly below those for "Gora", and the separation appears to be significant, although the sample sizes are small for both the works.

Table 8.7: Average sentence-length (no. of words) by decile groups based on probability samples of sentences from "Gora" and "Sheser Kavita".

decile group (per cent)	average of sentence-length (no. of words)					
	Gora			Sheser Kavita		
	h.s. 1	h.s. 2	comb.	h.s. 1	h.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0 - 10	3.89	3.93	3.91	1.86	3.58	2.44
10 - 20	6.78	6.14	6.34	3.45	5.16	4.12
20 - 30	8.44	8.27	8.42	4.32	6.58	5.55
30 - 40	9.66	9.68	9.63	6.30	8.26	7.38
40 - 50	11.22	11.62	11.43	8.03	9.26	8.78
50 - 60	12.33	13.73	13.11	10.07	10.42	10.32
60 - 70	14.66	15.44	15.05	11.90	12.53	12.13
70 - 80	17.33	18.90	18.18	13.74	14.84	14.36
80 - 90	21.88	26.60	24.49	16.90	17.74	17.28
90 - 100	39.21	49.07	44.84	30.09	25.32	27.75
0 - 100	14.54	16.34	15.54	10.67	11.37	11.02
no. of sample sentences	90	112	202	97	95	192

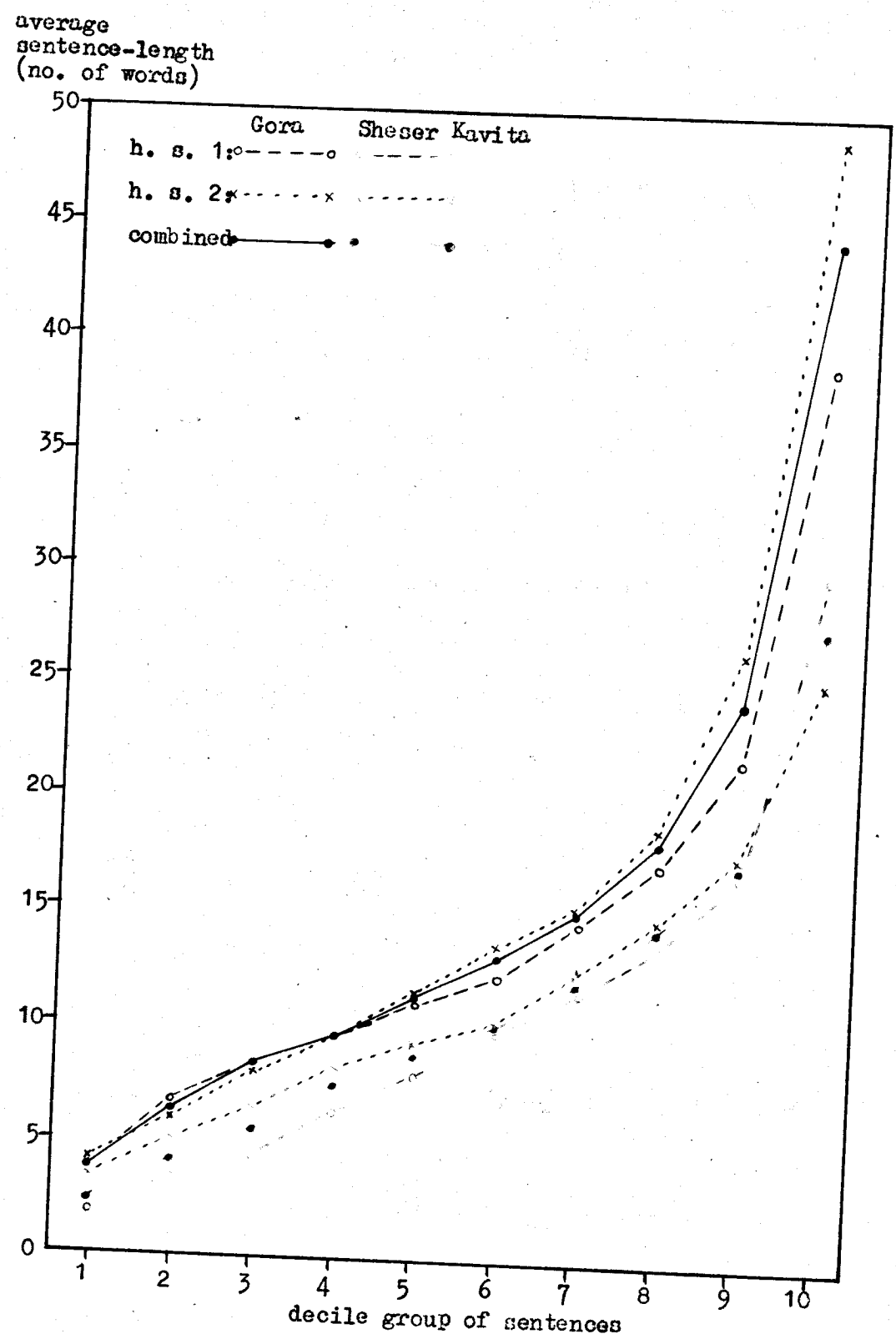


Fig. 8.2: Fractile graphs for sentence-length in terms of word based on probability samples of 202 and 192 sentences from "Gora" and "Sheser Kavita" respectively [Vide Table 8.7].

8.6.5. In the same way, the difference between "Durgeshnandini" and "Visavriksha", seems to be significant, but the difference is smaller here, the averages being 10.89 and 8.98 respectively. "Krishnakanter Will" (average = 9.02) has very nearly the same average as "Visavriksha"; and using the subsample averages for either the systematic sample or the probability sample, "Krishnakanter Will" seems to have a significantly lower average than "Durgeshnandini".

8.6.6. "Pather Panchali" (average = 11.60) has an appreciably larger average than "Devayan" (average = 8.90); and this is borne out significant by the subsample estimates and fractile graphs.

8.6.7. The other within author differences in average do not seem to be significant, excepting that "Birbaler Halkhata" has a significantly larger average (15.00) than "Char-Yari Katha" (11.15) but the first-named work is a collection of essays, while the second one is of the nature of fiction.

8.6.8. It is interesting to recall in this connection the findings reported in Section 3.4 of Chapter 3. A similar negative conclusion was reached there on the use of word-length as indicator of an author's style. The remarks made there regarding word-length as style-indicator apply to a great extent to sentence-length as a measure of style. But there is one important difference. Sentence-length does not seem to have any trend over time as shown by word-length during the last century.

8.6.9. Works showing appreciable within author differences in average sentence-length tend to show appreciable differences in the average length of words also. Consider, for example, "Birbaler Halkhata" and "Char-Yari Katha", or "Pather Panchali" and "Devayan". But there are exceptions to this. Thus, "Chaturanga" does not differ from "Gora" appreciably or significantly in the average of word-length; on the other hand, "Visavriksha" and "Krishnakanter Will" do differ appreciably and significantly in respect of the average of word-length.

8.6.10. There are some within author differences in respect of $\hat{\lambda}$ i.e., of C.V. also; and a few of these are in unexpected quarters. Thus, "Pallisamaj" and "Pather Dabi" are apparently very similar. They are found similar in respect of the average of sentence-length, and they were found similar in regard to word-length also. But the former work seems to have an appreciably smaller C.V. of sentence-length, although the difference may not be strictly significant. This may be due to "Pather Dabi" having much of elevated discourse on socio political topics in its conversations. The differences between "Sheser Kavita" and "Yogayog", or "Visavriksha" and "Krishnakanter Will", or "Deshe Videshe" and "Chacha Kahini" may also be mentioned here. These differences may not be significant in the strict sense, but taken together they show that C.V. of sentence-length may vary between apparently very similar works. There is a striking difference between "Birbaler Halkhata" and "Char-Yari Katha". The former has a much smaller C.V. But the two works are very different in nature, as already stated.

8.6.11. Yule (1938) claimed to have demonstrated that sentence-length is an indicator of style in prose, but this did not imply that different works of the same author would show the same type of sentence-length distribution. Yule recognised that sentence-length may be affected by subject-matter : argumentative passages may have longer sentences than purely descriptive ones (Yule, 1938, p.367).1/ Yule merely demonstrated the consistency of sentence-length distributions within similar works by a given author, mostly within the same collection of essays. Subba Rao (1960) sets out to establish that sentence-length is an indicator of an author's style, but omits to make any examination of within author differences. This line of study seems to merit further investigation. Williams (1940, p.361) was right in emphasising that a large number of within author comparisons are needed before one can generalise about the effects of (1) subject-matter, (2) age of author and (3) style variations within the same author; and until this is done, it would be unsafe to use sentence-length data for what has been called the "fingerprinting of authorship".

8.7.1. Lognormality of sentence-length distributions in Bengali : Lognormality of sentence-length was discovered by Williams (1940), who examined some works by three authors in English prose. Apart from the small scale of this investigation, Williams did not employ any objective

1/ For applications to the two cases of disputed authorship, Yule took care to compare works by the rival claimants which were similar in subject-matter to the work whose authorship was under dispute.

measure or test of goodness of fit to support his conclusion.[1] It is not meant to undermine the value of this pioneering attempt here; but it must be pointed out that the lognormality of sentence-length seems to be accepted on all hands without rigorous examination of sufficiently varied material. Thus, Subba Rao (1960) in his study on Kannada prose tacitly assumes lognormality, without any attempt to test it with his material.

8.7.2. Below we present the findings of a detailed examination of the lognormality of sentence-length distributions for the twenty works in Bengali prose covered in Table 8.2. Tables 8.3 to 8.5 present the relevant frequency distributions. The twenty works were studied separately; but in each case, only the overall distribution was examined, and not the separate distributions by type of sample or subsamples. In general, the deviations from unrestricted random sampling were ignored in this investigation; such deviations are expected to be small in view of the findings of Chapter 9.

8.7.3. The method of quantiles was used for estimating the parameters θ and λ of the underlying lognormal distribution. The two values of sentence-length (x) having the cumulative percentages of sentences nearest to 27% and 73% respectively were utilised for estimating θ,

---

1/ Curiously, Williams considered the grouped frequency distributions given by Yule (1938) as unsuitable for examination of lognormality, and found it necessary to prepare ungrouped frequency distributions for himself for carrying out such examinations (Williams, 1940, p.356).

and the two values with the cumulative percentages nearest to 7% and 93% respectively were used for estimating λ. This was done to maximise the efficiency of the estimates. (Vide Aitchison & Brown, 1957, pp. 40-42 .) The reasons for adopting this method of estimation will be clear from Section 6.5 of Chapter 6, if it is remembered that the observed sentence-length distributions are far less discrete than the observed **word** -length distributions.

8.7.4. The estimates of $\theta = E(\log_{10} x)$ and $\lambda = \sigma(\log_{10} x)$ are presented in cols. (11) and (12) of Table 8.2. The estimates of λ have already been used for drawing inferences. The observed and expected distributions of sentence-length are presented in Table 8.8 in the cumulative form.

8.7.5. Figure 8.3(a)-(c) presents the observed ogives on log-probit scale saparately for three works representing different types of situations. The subsamples of probability and systematic samples were merged to give two independent and interpenetrating half-samples. The observed points are given separately for these half-samples and also for the pooled sample. This enables one to examine the significance of the non-linearity, if any, in the observed ogive in a semi-intuitive manner.[1] It can be seen that the ogive is practically straight for "Krishnakanter Will", excepting for a trace of convexity at the very lowest value of sentence-length, viz. 1; the tendency is slightly clearer for "Kavi Shri-Ramakrishna" but quite clear for "Yogayog".

[1] This introduction of IPNS in the well-known log-probit diagram gives a more generally applicable tool than Linder's fractile graphical method which requires individual ungrouped observations (Linder, 1963).

Table 8.8: Cumulative distributions of sentence-length in twenty works in Bengali prose, along with expected distributions based on fitted lognormal curves.

(a) 16 works subjected to probability sampling only

length (no.of words)	cumulative percentages of sentences							
	Shakuntala		Sitar Vanavas		Durgeshnandini		Visavriksha	
	obgerved	expected	obgerved	expected	obgerved	expected	obgerved	expected
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1		$0.0^3 2$		$0.0^3 2$	1.35	0.72		0.11
2		0.04	0.36	0.03	3.37	2.04	3.05	2.39
3	1.36	0.47	0.72	0.34	8.08	7.13	10.37	8.96
4	2.27	1.96	1.81	1.41	15.82	14.56	18.90	18.60
5	6.36	4.69	2.54	3.57	23.91	23.06	28.66	29.36
6	7.73	8.92	5.80	6.93	30.30	31.69	36.59	39.85
7	11.82	14.31	9.78	11.32	39.06	39.89	46.34	49.37
8	17.73	20.46	14.49	16.50	45.12	47.39	56.71	57.67
9	28.64	27.02	22.10	22.16	53.20	54.09	64.63	64.72
10	34.55	33.67	26.81	28.06	58.92	60.00	71.34	70.64
11	40.45	40.16	33.70	33.98	64.31	65.15	77.44	75.56
12	47.27	46.34	42.03	39.75	68.35	69.63	79.27	79.64
13	51.82	52.11	47.10	45.28	72.73	73.50	81.71	83.00
14	56.82	57.41	50.00	50.47	77.10	76.85	86.59	85.77
15	63.18	62.24	55.07	55.31	79.46	79.74	87.80	88.07
16	67.27	66.58	59.78	59.76	82.49	82.24	89.63	89.96
17	69.09	70.47	63.77	63.83	84.85	84.40	92.07	91.53
18	72.27	73.94	68.48	67.53	85.19	86.28	93.90	92.84
19	76.82	77.01	72.10	70.88	88.55	87.90		
20	79.55	79.73	74.64	73.90	89.56	89.31	94.51	94.83
21	81.36	82.13	78.62	76.62	91.92	90.54	95.73	95.59
22	85.45	84.24	81.16	79.06	92.59	91.61	96.34	96.23
23	87.27	86.10	84.06	81.24	93.60	92.55	98.17	96.77
24	89.55	87.73	86.23	83.19	94.95	93.37	98.78	97.22
25	90.45	89.17	88.77	84.93	95.29	94.09	99.39	97.61
26	92.27	90.43			95.62	94.72	100.00	97.93
27	93.64	91.54			96.30	95.27		
28	94.09	92.52	89.49	89.13	96.63	95.76		
29	95.55	93.37	90.22	90.24	96.97	96.20		
30	96.36	94.12	91.30	91.23	97.64	96.58		
31	96.82	94.79	92.03	92.11	98.32	96.92		
33	97.27	95.89	92.39	93.61				
34			93.12	94.24	98.99	97.73		
35			93.84	94.81				
36	97.73	97.10	94.57	95.31	99.33	98.14		
37	98.18	97.41	95.29	95.77				
38	98.64	97.69	96.01	96.18				
39			96.74	96.54				
40			97.46	96.87	99.66	98.72		
43			97.83	97.67				
48					100.00	99.37		
49			98.91	98.68				
53	99.55	99.53	99.28	99.08				
58	100.00	99.71	99.64	99.41				
68			100.00	99.74				
no. of sample sentences	220		276		297		164	

Table 8.8 : (contd.)

(a) 16 works subjected to probability sampling only

length (no.of words)	cumulative percentages of sentences							
	Gora		Sheser Kavita		Birbaler Halkhata		Char-Yari Katha	
	observed	expected	observed	expected	observed	expected	observed	expected
(1)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1		0.03	1.04	0.25		0.04	4.52	0.44
2	0.50	0.71	5.21	2.87	0.46	0.02	6.03	3.70
3	3.47	2.93	9.37	8.59	1.39	0.30	11.56	9.70
4	6.93	6.77	18.75	16.18	2.31	1.36	16.08	17.04
5	11.39	11.81	24.48	24.41	4.63	3.69	23.62	24.69
6	15.35	17.56	30.21	32.52	8.80	7.44	27.64	32.10
7	19.80	23.61	35.94	40.13	12.50	12.43	36.18	38.99
8	25.74	29.66	42.71	47.05	21.30	18.35	45.23	45.26
9	33.66	35.53	49.48	53.23	27.31	24.82	52.76	50.91
10	40.10	41.09	56.77	58.70	36.57	31.50	58.29	55.94
11	45.54	46.28	62.50	63.51	39.35	38.13	63.32	60.41
12	52.48	51.08	66.15	67.73	43.98	44.50	68.34	64.38
13	56.44	55.48	71.35	71.41	47.22	50.50	72.36	67.90
14	62.87	59.50	75.00	74.64	53.24	56.05	75.38	71.03
15	66.83	63.16	80.73	77.46	59.26	61.11	76.88	73.80
16	69.80	66.47	83.85	79.93	62.50	65.68	78.39	76.26
17	73.27	69.47	85.42	82.09	67.13	69.78	82.41	78.46
18	76.24	72.18	88.02	83.99	69.44	73.42	83.42	80.41
19	78.71	74.64	89.06	85.66	74.07	76.65	85.93	82.16
20	80.20	76.86	90.10	87.14	79.17	79.51	89.45	83.72
21	82.18	78.86	91.15	88.44	80.56	82.02	90.45	85.13
22	83.17	80.68	93.75	89.59	85.65	84.22	91.96	86.39
23	84.65	82.32			87.96	86.16		
24	85.15	83.81	94.79	91.52	89.81	87.85	92.46	88.54
25	86.63	85.15	95.83	92.32	91.67	89.33	93.97	89.46
26	87.62	86.37	96.87	93.04	92.59	90.63	94.97	90.30
27	88.12	87.48			93.06	91.77	95.98	91.06
28	88.61	88.49			95.83	92.76	96.99	91.74
29	89.11	89.41			96.30	93.63		
30	89.60	90.24			97.69	94.39	97.99	92.93
31	91.09	91.00			98.15	95.05	98.49	93.45
32	92.08	91.69	97.40	96.01				
33	92.57	92.33	98.96	96.35				
34	93.07	92.90						
36	93.56	93.92						
37	94.06	94.36					99.00	95.45
39	94.55	95.14						
40	95.05	95.48			98.61	98.13	99.50	96.30
41								
42	95.54	96.09			99.07	98.52		
44	96.04	96.60			99.54	98.69		
46	97.03	97.04						
47								
48	97.52	97.42	99.48	98.80				
49			100.00	98.97				
50	98.02	97.74						
51	98.51	97.88			100.00	99.69		
60	99.01	98.79					100.00	98.25
75	99.50	99.48						
93	100.00	99.79						
no. of sample sentences	202		192		216		199	

Table 8.8 : (contd.)

(a) 16 works subjected to probability sampling only

length (no.of words)	cumulative percentages of sentences							
	Palligamaj		Pather Dabi		Pather Panchali		Devayan	
	observed	expected	observed	expected	observed	expected	observed	expected
(1)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)
1	0.90	0.03	1.97	0.18	2.03	0.09	1.22	0.36
2	2.69	0.75	4.93	2.08	2.54	1.68	6.08	4.41
3	5.27	3.40	8.87	6.41	5.08	6.21	15.20	12.89
4	10.65	8.15	12.81	12.41	11.17	13.11	24.92	23.42
5	18.72	14.42	18.72	19.19	18.27	21.24	33.13	34.10
6	23.66	21.49	25.62	26.15	30.46	29.66	43.77	43.96
7	29.49	28.80	33.00	32.90	39.09	37.79	50.76	52.62
8	37.56	35.96	37.93	39.25	45.18	45.32	59.88	60.05
9	44.28	42.72	44.83	45.11	52.28	52.11	64.74	66.31
10	50.11	48.95	47.78	50.45	58.88	58.13	69.00	71.56
11	57.29	54.63	55.67	55.27	62.44	63.43	74.47	75.95
12	59.98	59.72	62.07	59.61	67.51	68.05	79.94	79.60
13	63.57	64.27	65.02	63.50	71.07	72.07	83.28	82.66
14	67.60	68.32	67.49	66.99	74.62	75.56	84.80	85.21
15	71.19	71.89	71.92	70.10	78.68	78.59	87.84	87.35
16	75.23	75.05	73.40	72.89	82.23	81.21	90.58	89.14
17	80.16	77.84	74.88	75.38	84.26	83.48	93.01	90.66
18	81.95	80.29	76.85	77.62	86.29	85.45	93.92	91.93
19	85.09	82.45	80.30	79.62	87.81	87.17	94.53	93.02
20	86.89	84.36	82.27	81.42	88.83	88.66	95.14	93.94
21	89.13	86.05	85.71	83.03	89.85	89.96	96.05	94.72
22	90.47	87.53	87.19	84.49	90.36	91.09		
23	92.72	88.84	88.18	85.80			96.66	95.97
24	93.16	90.01	89.16	86.98	91.88	92.96	96.96	96.46
25			90.64	88.05	92.39	93.72	97.57	96.89
26	95.41	91.95	91.63	89.01	94.42	94.39	98.48	97.26
27	95.85	92.76	92.61	89.89	94.92	94.98		
28			94.09	90.68				
29	96.30	94.12	95.07	91.41	95.94	95.96		
30	97.20	94.69	95.57	92.06	96.45	96.37		
31			96.06	92.66	97.46	96.73		
32					97.97	97.05		
33	97.65	96.06					98.78	98.80
34	98.54	96.42	97.54	94.16	98.48	97.60		
35					98.98	97.82		
36	98.99	97.05	98.03	94.96			99.39	99.14
38			98.52	95.64				
39	100.00	97.76	99.51	95.94			99.70	99.37
45			100.00	97.29				
50							100.00	99.78
56					99.49	99.65		
78					100.00	99.93		
no. of sample sentences	223		203		197		329	

Table 8.8 : (contd.)

(a) 16 works subjected to probability sampling only

length (no.of words)	cumulative percentages of sentences							
	Dristipat		Janantik		Chacha-Kahini		Deshe-Videshe	
	observed	expected	observed	expected	observed	expected	observed	expected
(1)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	(33)
1	1.91	0.37	2.07	0.55	1.75	0.28	0.53	0.15
2	5.74	4.18	7.44	5.00	5.26	2.81	2.12	2.13
3	11.96	11.97	14.05	13.18	10.09	7.98	7.94	7.01
4	21.05	21.67	19.83	22.83	17.54	14.69	16.40	13.95
5	30.62	31.61	27.69	32.47	23.68	21.96	24.87	21.83
6	35.89	40.92	39.26	41.37	29.82	29.21	30.69	29.84
7	44.98	49.24	50.41	49.29	35.53	36.08	35.45	37.50
8	53.11	56.49	58.68	56.18	40.79	42.44	41.80	44.58
9	60.29	62.71	67.77	62.11	47.37	48.22	47.09	50.97
10	67.46	68.02	71.90	67.20	51.75	53.43	52.38	56.68
11	73.68	72.52	75.62	71.54	56.58	58.09	56.08	61.73
12	80.38	76.34	78.51	75.24	59.21	62.25	61.38	66.16
13	83.73	79.58	81.82	78.41	64.04	65.95	63.49	70.07
14	84.21	82.33	85.12	81.12	67.98	69.26	68.78	73.48
15	85.65	84.67	87.19	83.44	70.18	72.19	73.55	76.47
16	88.52	86.66	88.84	85.44	74.12	74.82	76.19	79.09
17	89.95	88.36	90.91	87.17	75.88	77.15	79.89	81.39
18	91.87	89.82	94.63	88.66	76.75	79.24	83.07	83.41
19	93.30	91.07			79.39	81.10	85.71	85.18
20	93.78	92.15	95.87	91.08	80.26	82.77	87.83	86.74
21	94.74	93.08	96.28	92.06	83.33	84.27	90.48	88.12
22					85.09	85.62	91.53	89.33
23			97.11	93.66	87.28	86.83	92.59	90.41
24	96.17	95.20	97.93	94.32	90.35	87.93		
25	96.65	95.73	98.76	94.90	90.79	88.91	93.12	92.20
26	97.13	96.19			92.11	89.81	94.18	92.95
27			99.17	95.86	93.86	90.61	95.24	93.62
28	97.61	96.96			95.61	91.35	96.30	94.22
29	98.09	97.27					96.83	94.75
30	98.56	97.55	99.59	96.94				
31					96.05	93.17		
33	99.04	98.21			96.49	94.13		
34					96.93	94.56		
36	99.52	98.67					97.35	97.24
37	100.00	98.80			97.37	95.62		
38					98.25	95.92		
39					99.12	96.19		
40					99.56	96.44	97.88	98.03
42			100.00	98.95				
50					100.00	98.13	98.94	99.10
58							99.47	99.49
93							100.00	99.93
no. of sample sentences	209		242		228		189	

Table 8.8: (contd.)

(b) 3 works subjected to both probability and systematic sampling

length (no.of words)	cumulative percentages of sentences					
	Krishnakanter Will		Yogayog		Kavi Shri Ramakrishna	
	observed	expected	observed	expected	observed	expected
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1.45	0.48	0.68	0.05	0.47	0.17
2	5.79	5.03	3.25	1.35	4.47	3.81
3	14.62	13.87	8.38	5.60	14.24	13.64
4	24.31	24.45	14.70	12.55	26.47	26.36
5	34.01	34.98	21.88	21.01	40.71	40.31
6	44.72	44.59	29.40	29.94	52.47	52.34
7	52.53	53.00	37.95	38.59	63.41	62.38
8	58.90	60.19	46.67	46.61	70.82	70.47
9	64.69	66.26	53.85	53.78	77.29	76.86
10	71.49	71.37	59.15	60.14	82.82	81.86
11	76.27	75.64	65.30	65.65	87.06	85.74
12	80.32	79.22	70.94	70.40	90.35	88.76
13	82.92	82.22	76.41	74.50	92.00	91.11
14	86.25	84.74	79.32	77.99	93.88	92.93
15	88.71	86.86	83.08	81.00	95.88	94.36
16	89.87	88.66	87.01	83.57	97.06	95.48
17	91.17	90.17	89.23	85.77	97.53	96.36
18	92.48	91.46	91.11	87.66	98.00	97.06
19	93.49	92.56	92.65	89.27	98.24	97.61
20	94.36	93.50	94.19	90.64	98.59	98.06
21	95.08	94.31	95.04	91.83	98.82	98.41
22	95.66	95.00	95.73	92.65	99.18	98.69
23	96.24	95.60	96.41	93.74	99.29	98.92
24	96.67	96.11	96.75	94.51	99.53	99.11
25	96.82	96.56	96.92	95.16	99.65	99.26
26	97.11	96.95	97.61	95.75		
27	97.68	97.29	97.78	96.24	99.76	99.49
28	97.83	97.59	98.46	96.67		
29	98.12	97.85	98.63	97.06	99.88	99.64
30			98.80	97.39		
31	98.26	98.28	98.97	97.68	100.00	99.74
32			99.15	97.93		
33	98.41	98.61	99.49	98.16		
34	98.70	98.75		98.35		
37			99.66	98.82		
38	98.84	99.17				
39	98.99	99.24	99.83	99.04		
40	99.13	99.31				
41	99.42	99.38				
42	99.57	99.43				
43	99.71	99.48				
50			100.00	99.68		
51	99.86	99.74				
72	100.00	99.95				
no. of sample sentences	691		585		850	

Table 8.8: (contd.)

(c) "Chaturanga", Parts 1-2, subjected to complete count.

length (no. of words)	cumulative percentage of sentences	
	observed	expected
(1)	(2)	(3)
1	0.30	0.04
2	1.63	0.93
3	4.89	3.98
4	11.41	9.21
5	17.93	15.92
6	24.89	23.32
7	31.11	30.84
8	37.78	38.09
9	43.56	44.87
10	50.81	51.07
11	56.59	56.66
12	60.30	61.65
13	63.41	66.08
14	67.41	69.99
15	72.00	73.44
16	75.85	76.47
17	78.07	79.14
18	81.48	81.48
19	84.30	83.54
20	87.11	85.35
21	88.89	86.94
22	90.07	88.34
23	91.56	89.58
24	92.44	90.67
25	93.04	91.64
26	93.78	92.50
27	95.11	93.26
28	96.15	93.93
29	96.30	94.53
30	96.89	95.06
31	97.48	95.54
32	97.93	95.96
33	98.52	96.34
35	98.67	96.99
36	98.81	97.26
37	98.96	97.51
38	99.11	97.73
39	99.26	97.93
41	99.41	98.27
45	99.56	98.78
51	99.70	99.26
63	99.85	99.70
100	100.00	99.97
no. of sentences	675	

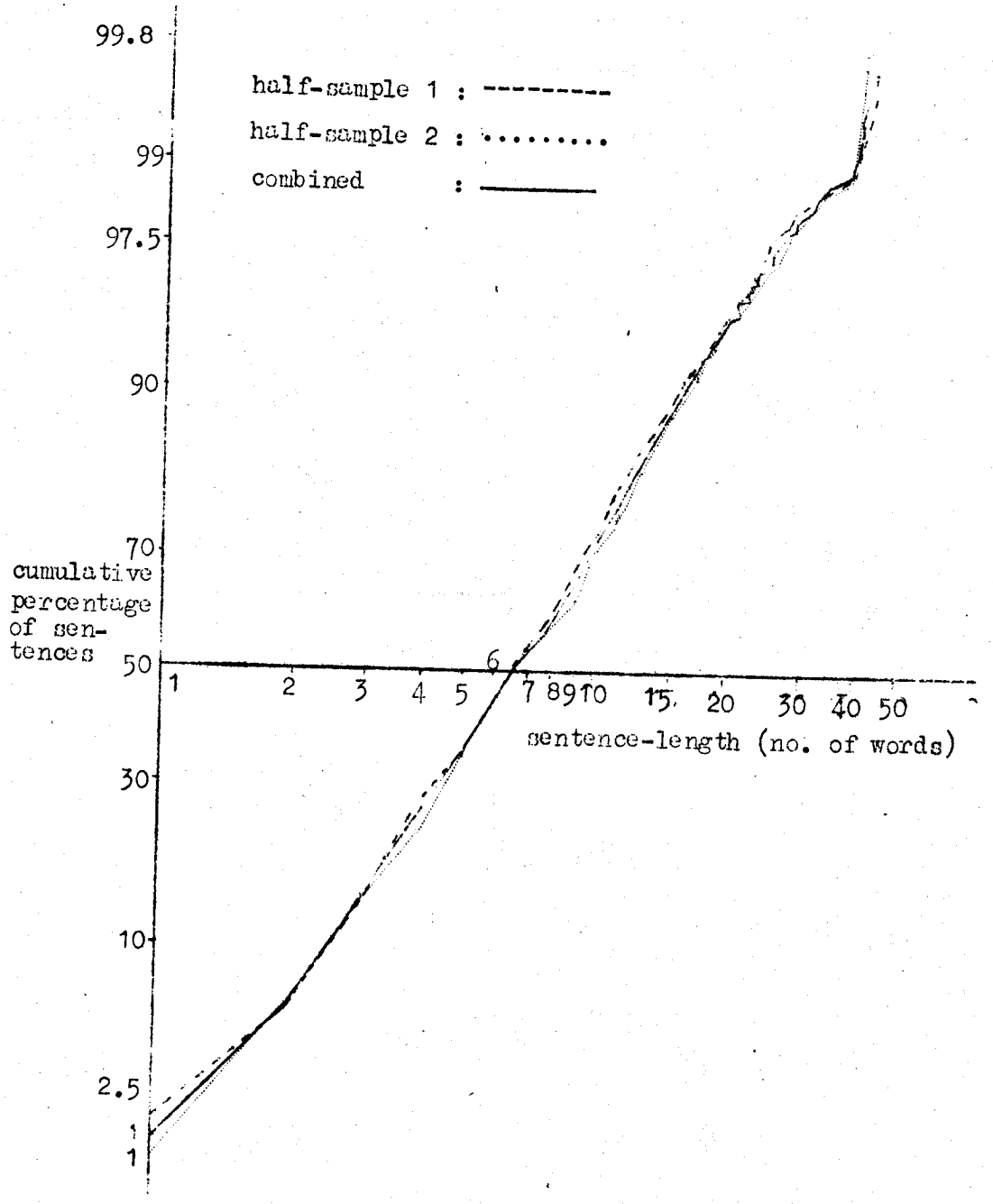


Fig.8.3(a): Ogive, on log-probit scale, for the distribution of sentences by length in words, based on the pooled (probability plus systematic) sample of 691 sentences from "Krishnakanter Will", separately for the pooled sample and for its two half-samples.

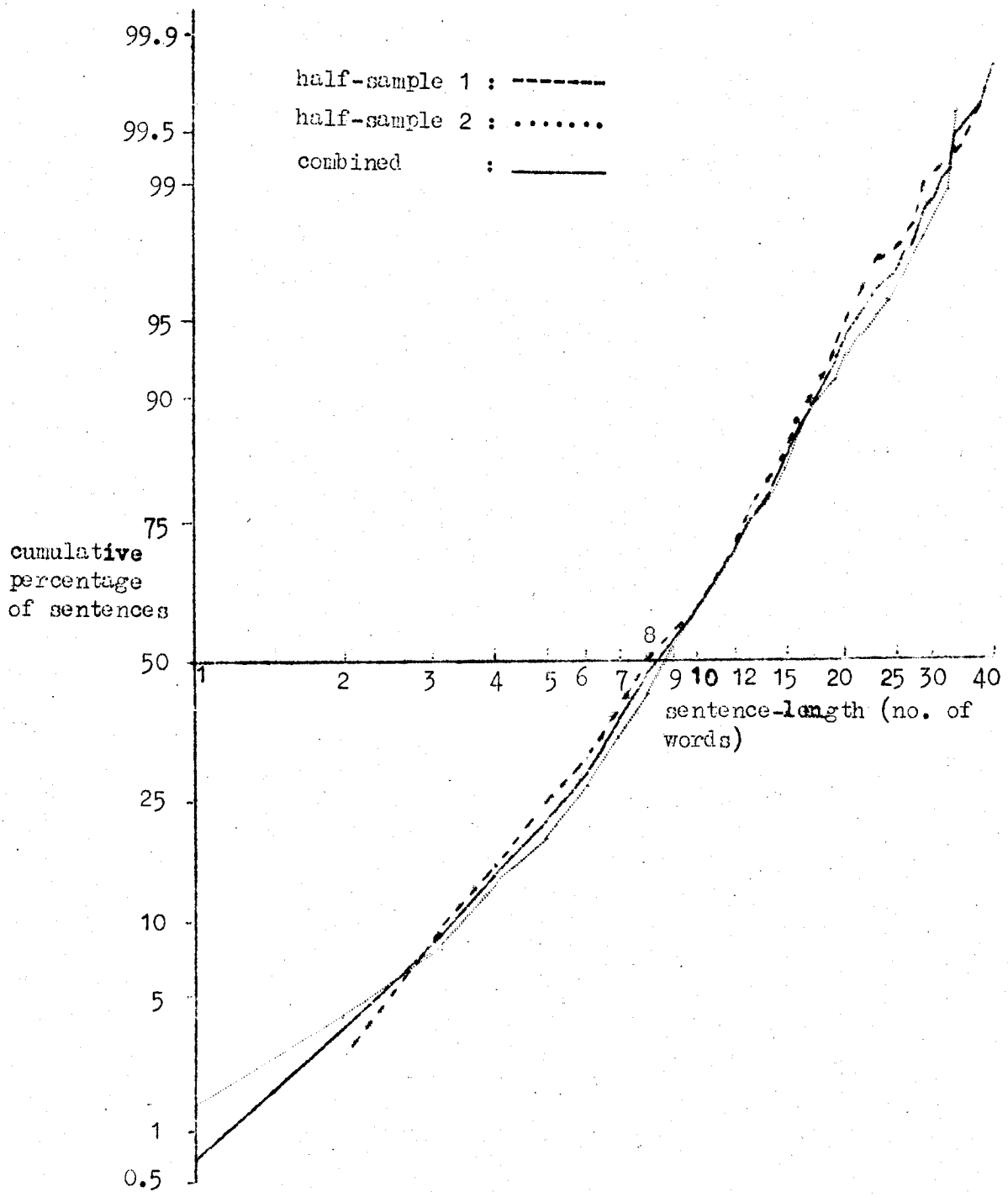


Fig.8.3(b): Ogive, on log-probit scale, for the distribution of sentences by length in words, based on the pooled (probability plus systematic) sample of 585 sentences from "Yogayog", separately for the pooled sample and for its two half-samples.

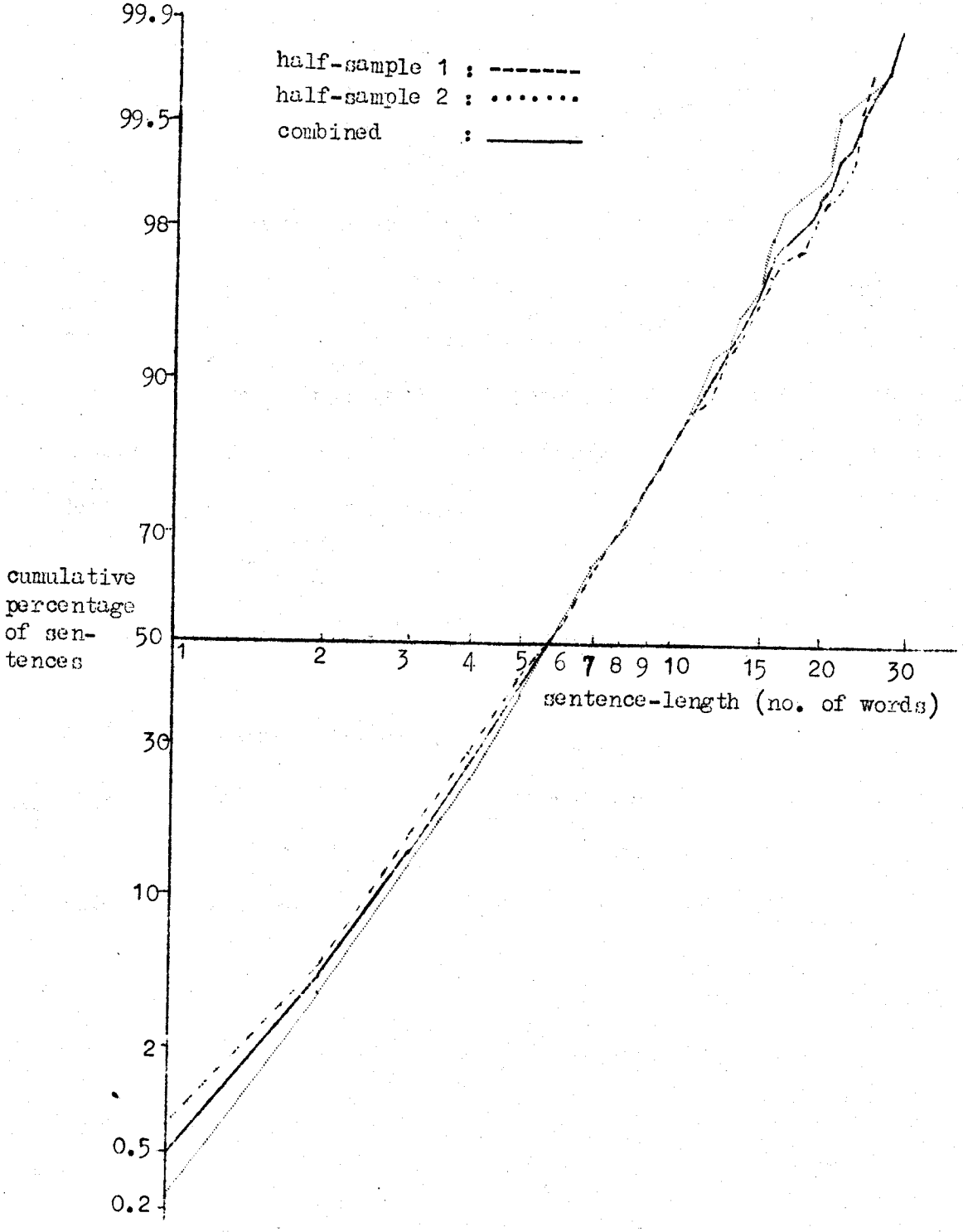


Fig.8.3(c): Ogive on log-probit scale, for the distribution of sentences by length in words, based on the pooled (probability plus systematic) sample of 850 sentences from "Kavi Shri Ramakrishna", separately for the pooled sample and for its two half-samples.

8.7.6. Table 8.9 presents the results of the Kolmogorov test applied for examining the goodness of fit of the lognormal distribution. The observed distribution is discrete; and the sample is not unrestricted random but has some elements of clustering. These two difficulties do not seem to be serious. Sentence-length is nearly continuous; and in view of the finding in Chapter 9, the samples of sentences are nearly random. What is more serious is that the expected distribution is obtained after estimating two parameters from the observed data. This should make the K-test conservative in the present situation (vide Section 6.6 Chapter 6, for a detailed discussion).

8.7.7. Not a single work in Table 8.9 shows a significant value of K; indeed no value reaches the (upper) 20% point of the distribution. Denoting by n the number of sample sentences, the values of $\sqrt{n}K$ are distributed over the range 0.4 to 0.9, roughly speaking, while the upper 20% point is 1.07 for large n. Apparently, the K-test gives non-significant results, on the whole, also. The K-values range from about 1.5% to 6.5%, and tend to decrease with sample size.

Table 8.9: Examination of lognormality of the sentence-length distributions for twenty works in Bengali prose, presented in Tables 8.3 to 8.5[1]

works	type of sample	no. of sample sentences	Kolmogorov distance[2] K(%)	remarks re: linearity of ogive on log-probit scale
(1)	(2)	(3)	(4)	(5)
1. Shakuntala	prob.	220	2.73	slight convexity
2. Sitar Vanavas	"	276	3.84	-do- at the lower end
3. Durgeshnandini	"	297	2.27	moderate convexity
4. Visavriksha	"	164	3.26	slight convexity at the higher end
5. Krishnakanter Will	pooled	691	1.85	moderate convexity
6. Gora	prob.	202	4.07	nearly perfect linearity
7. Chaturanga (Parts I & II)	complete count	675	2.67	moderate convexity
8. Sheser Kavita	prob.	192	4.16	appreciable convexity
9. Yogayog	pooled	585	3.55	pronounced convexity
10. Birbaler Halkhata	prob.	216	5.07	appreciable convexity
11. Char-Yari Katha	"	199	5.73	pronounced convexity
12. Pallisamaj	"	223	4.30	moderate convexity
13. Pather Dabi	"	203	3.66	-do-
14. Pather Panchali	"	197	2.97	some convexity at lower end
15. Devayan	"	329	2.56	-do-
16. Dristipat	"	209	5.03	moderate convexity
17. Janantik	"	242	5.97	pronounced convexity
18. Chacha-Kahini	"	228	4.26	moderate convexity
19. Deshe-Videshe	"	189	6.58	wave-like deviations, but no convexity
20. Kavi Shri Ramakrishna	pooled	850	1.59	moderate convexity

1/ Vide Table 8.8(a)-(c) for detailed comparisons between observed and expected distributions.

2/ In all cases, the P-value is above 20%.

8.7.8. It might appear from the above that the lognormal hypothesis gives a good fit to the sentence-length distributions for Bengali prose. But Col. (5) of Table 8.9 points to some systematic divergence between observation and theory : The log-probit graphs are convex to the horizontal axis for most of the twenty distributions. [Vide Figs. 8.3(a)-(c)]. These graphs are a little misleading in the sense that the same vertical deviation between observed points and the fitted line would mean larger differences between the corresponding cumulative proportions in the middle of the range than near the two extremes. Nevertheless, the tendency towards convexity is unmistakable. The following counts are based on Table 8.8 :

sentence-length (x)	no. out of the twenty distributions where $F_o(x)$ exceeds $F_e(x)$
(1)	(2)
1	15
5	10
10	9
20	16
30	19
40	17

This shows the effect of convexity : both at the upper ranges of sentence-length and also at the very lowest values, $F_o(x)$ tends to be systematically larger than $F_e(x)$.

8.7.9. A partial explanation of this convexity seems to be the heterogeneity of the population of sentences. Some sentences are

conversational, being completely embedded within conversational matter; other sentences are completely outside conversational matter; and there is a third category of mixed sentences partly within conversations and partly outside. Table 8.10 indicates that conversational sentences are appreciably shorter, on the average, than sentences of the other two categories. The 88 conversational sentences in "Chaturanga", Parts 1 and 2, have 9.92 words each, on the average; the average for 481 non-conversational sentences is 12.89 and that for 106 mixed sentences 12.41. Fig. 8.4 indicates that the nonconversational sentences (considered separately) follow the lognormal distribution fairly closely.1/ This may be true for all three classes of sentences. When a lognormal curve is fitted to a mixture of lognormal distributions with different means, it is but natural to find a little too many observations near the extremes and a little too few near the middle, thus getting the observed convex type of ogive on log-probit graph. This line of thought could not be pursued further, but the point seems to be deserve careful attention.

8.7.10. It may also be mentioned that even if the sentence-length distribution is exactly lognormal, defective punctuation may give rise to unnecessarily long sentences and produce a convex type of ogive on the log-probit diagram. This might be another partial explanation of the observed convexity.

1/ The distribution being based on a complete count, half-samplewise ogives could not be shown in Fig. 8.4.

Table 8.10: Sentence-length distributions based on a complete count of Parts 1 and 2 of Tagore's "Chaturanga", separately for three categories of sentences.

length (no.of words)	no. of sentences							
	Part 1				Part 2			
	conversational	non-conv.	mixed	total	conversational	non-conv.	mixed	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1					1	1		2
2	6			6	1	2		3
3	2	6	1	9	3	8	2	13
4	8	11	2	21	1	18	4	23
5	2	19	3	24	3	9	8	20
6	4	17	3	24	1	19	3	23
7	3	14	5	22	1	17	2	20
8	5	16	5	26	1	15	3	19
9	2	14	3	19	2	16	2	20
10	4	27	1	32	3	13	1	17
11	4	18	5	27	1	10	1	12
12		10	3	13	2	10		12
13	4	9	1	14	1	4	2	7
14	2	7	7	16	1	10		11
15	1	16	5	22	2	6	1	9
16	2	5	4	11	1	9	5	15
17	2	5	1	8		6	1	7
18	3	9	2	14	2	6	1	9
19	1	7		8		8	3	11
20		11	2	13		4	2	6
21	2	5	1	8		4		4
22		4		4		3	1	4
23	1	4	3	8		2		2
24	1	2		3		3		3
25		2	1	3			1	1
26		3		3		2		2
27		3	2	5		4		4
28	1	2	1	4		3		3
29		1		1				-
30		3		3		1		1
31	1	2		3		1		1
32						3		3
33		2	1	3		1		1
35						1		1
36		1		1				-
37		1		1				-
38							1	1
39						1		1
41						1		1
45						1		1
51		1		1		-		-
63						1		1
100						1		1
total	61	257	62	380	27	224	44	295
average length	10.31	12.85	13.23	12.51	9.04	12.93	11.25	12.32

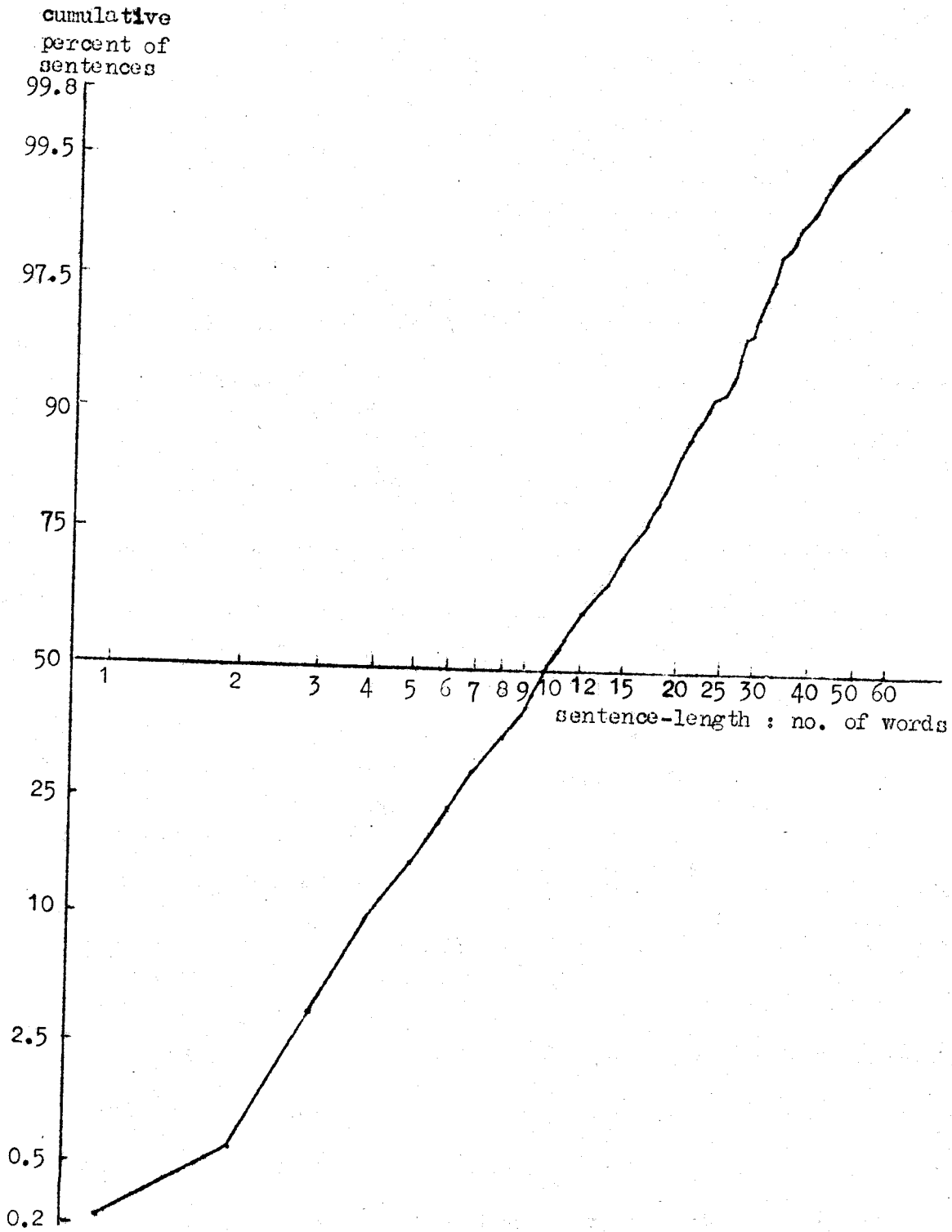


Fig. 8.4: Ogive, on log-probit scale, for the length-distribution of 481 wholly conversational sentences in "Chaturanga" (Parts 1-2) by Tagore.

Chapter 9: Sentence-length in Bengali Prose - (2)

9.1.1. Introduction : The present chapter is a continuation of Chapter 8, and reports on some further studies on sentence-length mostly confined to Bengali prose.

9.1.2. If the lengths of different sentences of a prose work are recorded in the natural reading order, one gets something like a time series which may be called the sentence-length series. The randomness of this series for some Bengali and English works will be examined in this chapter. Several types of nonparametric tests will be applied for such examination. The autocorrelation coefficients of different orders will be studied in this connection. It will be found that the series is nearly random with little trend or oscillations, and the autocorrelation coefficients of even the first few orders are of the order of 0.1, if not zero.

9.1.3. Chapter 5 on the randomness of word-length series may be referred to in this connection. The objective here is generally similar, but the scale of investigation is comparatively modest. The methods used are also somewhat different. Thus, the emphasis here is on testing the randomness of sentence-length in short continuous passages; the autocorrelations for complete works were not studied in any detail. A reasonable estimate of r_1, say, would require a fairly large sample of sentences. Selecting 1000 sentences, it will be appreciated, is much more laborious than selecting 1000 words; in fact, complete enumeration of

sentences may often be quicker. On the other hand, many tests of randomness are used here, which are not quite applicable to word-length series because of the highly discrete nature of the word-length distribution, leading to too many ties in any process of ranking.

9.1.4. The randomness of sentence-length series does not seem to have been examined so far and yet Williams (1940) and Subba Rao (1960) implicitly assumed that the series are approximately random.[1/] Yule (1938, p.371) remarks that the sentence-length series seems to be autocorrelated, and "short sentences tend to occur together". This, according to him, creates difficulties in sampling and in statistical inference.

9.1.5. Sections 9.2 to 9.4 are concerned with the non-parametric tests of autocorrelation etc., in selected sentence-length series from Bengali and English works. Section 9.5 discusses the systematic samples already used in the preceding chapter, and establishes their validity as approximations to probability samples of sentences. The corresponding studies for word-length may be found in Chapter 2, Section 2.6-7. Certain other studies are reported in Section 9.6 which point to the high degree of homogeneity or randomness of the series of sentence-lengths.

1/ Williams (1940) made this assumption in calculating standard errors of the estimates of lognormal parameters; and Subba Rao (1960) applied statistical methods valid for unrestricted random sampling. But both had used samples of sentences with a great deal of clustering involved in them.

9.2.1. Tests of Randomness : Tables 9.1(a) and (b) (to be explained later) present the results of several types of tests of randomness applied to each of a number of sentence-length series. The series were derived by complete counts of certain texts or extracts so that continuous series of sentence-lengths were obtained.

9.2.2. Actually, sentence-length series were obtained from the following texts or extracts :

(i) First two out of the four parts of Tagore's novel "Chaturanga";

(ii) an extract from Tagore's novel, "Sheser Kavita", namely the whole of the chapter entitled "Lavanya Punavritta" which hardly involves any conversational matter;

(iii) the three short essays covered for word-length and sentence-length studies, viz., "Samya" by Bamkimchandra, and "Bamkimchandra" and "Vishwavidyalay" by Tagore; and

(iv) two extracts from Jane Austen's "Pride and Prejudice", one from Chapter 35 and the second from Chapters 43-44.

9.2.3. The one-way distributions and averages have been presented and discussed in Chapter 8 (vide Table 8.5). The attention is now turned to the randomness of the sentence-length series produced from the above texts or extracts.

Table 9.1: Results of randomness tests on certain series of sentence lengths
(a) tests based on runs up and down and on Kendall's rank correlation coefficient τ between the serial order and the averages of subgroups

sr. no.	source text	no. of sentences	total no. of turning points (p) observed	expected	critical ratio	observed (expected) no. of phases length 1	length 2	length 3 or more	Wallis-Moore χ^2	test based on τ no. of sub-groups	no. of observations per sub-group	observed τ	critical ratio
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Bengali fiction												
1.	Chaturanga : Pt.I	380	241	252	-1.434	142 (157.08)	67 (68.93)	31 (24.98)	2.952	38	10	-0.178	-1.559
2.	-do- : Pt.II	295	199	195.3	0.300	125 (121.67)	52 (53.35)	21 (19.32)	0.272	29	10	-0.025	-0.169
3.	-do- :combined	675	442	448.6	-0.571	270 (280)	119 (123.02)	52 (44.65)	1.698	67	10	-0.086	-1.028
4.	Sheser Kavita:extract	155	102	102	0	62 (63.33)	29 (27.68)	10 (9.98)	0.091	31	5	-0.260	-2.040*
	Bengali essays												
5.	Samya	179	123	118	0.802	80 (73.33)	32 (32.08)	9 (11.58)	1.182	35	5	-0.039	-0.312
6.	Bamkimchandra	139	90	91.3	-0.168	53 (56.67)	22 (24.75)	12 (8.92)	1.609	27	5	0.208 1/	1.501
7.	Vishwavidyalay	125	80	82	-0.321	52 (50.83)	15 (22.18)	12 (7.98)	4.374	25	5	0.093	0.631
	English fiction												
8.	Pride and Prejudice:extract I	106	74	69.3	0.968	50 (42.92)	18 (18.70)	5 (6.72)	1.634	35	3	0.066	0.539
9.	-do- :extract II	189	128	124.6	0.608	79 (77.08)	33 (33.37)	10 (12.18)	0.979	37	5	0.102	0.876

* denotes significance at 5% level.

1/ If the 27th observation is excluded, τ falls to 0.145.

Table 9.1 : (contd.)

(b) Wald-Wolfowitz tests of significance of circular autocorrelation coefficients

sr. no.	source text	no. of sentences	circular autocorrelation coefficients r_1	r_2	r_3	r_6	r_{10}	common standard error	critical ratios for different coefficients t_{r_1}	t_{r_2}	t_{r_3}	t_{r_6}	$t_{r_{10}}$
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Bengali fiction												
1.	Chaturanga : Pt.I	380	0.082	0.127	0.044	-0.009	-0.014	0.0510	1.664	2.535*	0.922	-0.127	-0.216
2.	-do- : Pt.II	295	0.177	0.155	0.120	0.107	0.051	0.0556	3.244**	2.851**	2.220*	1.978*	0.980
3.	-do- :combined	675	0.137	0.145	0.088	0.059	0.017	0.0379	3.618***	3.824***	2.333*	1.557	0.451
4.	Sheser Kavita :extract	155	0.099	-0.041	-0.071	-0.094	0.082	0.0789	1.341	-0.439	-0.822	-1.112	1.125
	Bengali essays												
5.	Samya	179	-0.035	0.124	0.037	-0.012	-0.073	0.0741	-0.391	1.753	0.572	-0.079	-0.911
6.	Bamkimchandra	139	0.162 1/	-0.033	-0.008	0.097	0.172	0.0815	2.083*	-0.313	-0.013	1.275	2.197*
7.	Vishwavidyalay	125	-0.027	0.035	-0.069	0.070	-0.057	0.0872	-0.220	0.496	-0.701	0.896	-0.559
	English fiction												
8.	Pride and Prejudice :extract I	106	-0.102	-0.039	0.094	-0.197	-0.082	0.0944	-0.976	-0.310	1.099	-1.832	-0.773
9.	-do- : extract II	189	0.166	0.046	0.054	0.153	-0.017	0.0725	2.311*	0.711	0.822	2.182*	-0.155

* denotes significance at 5% level, ** at 1% level and *** at 0.1% level

1/ If the last two sentences having 75 and 56 words are excluded, non-circular r_1 falls to 0.012

9.2.4. For purposes of illustration, a bivariate distribution is presented in Table 9.2 showing the joint distribution of lengths of two consecutive sentences in Parts 1 and 2 combined of Tagore's "Chaturanga". The series comprised 675 sentences, and 675 pairs of consecutive sentences were formed, including the 'circular' pair. Fig. 9.1(a) presents the same data in the form of a scatter diagram. Fig. 9.1(b) shows the scatter diagram for the series obtained from the second extract from "Pride and Prejudice", having 189 sentences.[1/] (Some pairs of values occurred more than once, but this could not be indicated in the scatter diagrams.) It is evident from Table 9.2 and and Figs. 9.1(a) and (b) that lengths of consecutive sentences are independent in the statistical sense, at least to a rough approximation.

9.2.5. The undermentioned distribution-free or nonparametric tests of randomness were applied to each series. (Randomness means that the different observations are independent samples from the same underlying population, so that they have the same mean, standard deviation etc. There would then be no trend or oscillation, or more generally, no systematic movement in the series.) All these tests are conditional tests, that is to say, they take the observations $x_1, x_2, \ldots, x_n$ as given, and consider the conditional distribution of the test criteria over the n! permutations of the observations, which are equally likely under the null hypothesis. [Vide Walsh (1962, Chapter 5), which proved to be extremely helpful in the choice of tests.]

1/ The corresponding joint distribution of sentence-lengths could not be presented because of the unwieldy nature of the table.

Table 9.2: Joint distribution of lengths of consecutive sentences in terms of number of words : All sentences in Parts 1 and 2 of "Chaturanga" by Tagore.

length of preceding sentence	length of following sentence																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1		1								1												
2		1	1	2	1					1	1	1		1								
3	1	1		4	4	1	1	2	1	4	1	1							1			
4		1	1	4	5	1	5	5	* 4	1	3	1			2	3		1	1	3		1
5	1	1	3	3	7	4	1	5	3	1		2	2				1	1		2	2	
6			3	3	4	1	4	3	6	2	3	3		2	3	1	2		1	1	1	
7			2	2		6	1	2	3	2	2	1	2	2	2	1	2	3	4			
8		1	1	4	1	3	3	5	3	2	3		2		2	4	2	2		1		
9			1	1	1	5	4	3	4	2	5	4		2	1	1			1	2	1	
10			1	1	5	6	5	1	6	5	4	1		1	2	3		1	1	1		1
11			2	2	2	2	1	1	2	3	4	2	1	4	4	2	1	1	1			2
12		1	1	1	3	1		6		3	1	1	1	1	1	1			1	1		
13			1	2			2	3	2	3						1	2	1	1			
14		1	1	1	1	1	2	2		5			3		3			2			1	
15			1	4	2	2	2	1		4	1	1	2		2	1			1	1	1	1
16			1	5	1	1	2	2			4	2				1	2	1			2	
17				1	1	3	1	1		2				1		1		1				
18				1	2	1	1		1	1	3	1	2	2	1	1		1	1	2	1	
19					1	2					1	1		3	1	1	2	2	1			
20							2						1		3	1		2	1			2
21						1			1	2	1			2	1			1				1
22				1	2					1	1		1					1				
23						1	2	1		1	1				1	1					1	
24										1				1				1	1	1		
25													1	2							1	
26						1			1	1							1		1			
27	1	1				1	1		1	1				1					1			
28				1		1		1	1			1				1		1				
29						1																
30												1			1					2		
31							1	1					1									
32																		1		1	1	
33			1			1	1						1									
35																1						
36					1																	
37													1									
38															1							
39				1																		
41														1								
45											1											
51																						
63				1																		
100														1								
total	2	9	22	44	44	47	42	45	39	49	39	25	21	27	31	26	15	23	19	19	12	8

* includes the circular pair

(contd.)

Table 9.2: (contd.)

length of preceding sentence	length of following sentence																						total
	23	24	25	26	27	28	29	30	31	32	33	35	36	37	38	39	41	45	51	63	100		
1																							2
2																							9
3																							22
4	1					1																	44
5			1		1			2	1														44
6		1	1				1												1				47
7			1		1			1	1	1													42
8		1			3	1			1														45
9					1																		39
10			1		1			1	1	1													49
11				1												1							39
12		1																					25
13	1			1							1			1									21
14	2	1									1												27
15						1							1				1	1					31
16	1					1																	26
17	1			1							1												15
18	1																						23
19				1	1	1				1													19
20	2				1						1	1			1								19
21						1															1		12
22		1																					8
23				1																			10
24						1																	6
25																							4
26																							5
27																				1			9
28																							7
29																							1
30																							4
31		1																					4
32																							3
33																							4
35																							1
36																							1
37																							1
38																							1
39																							1
41																							1
45																							1
51	1																						1
63																							1
100																							1
total	10	6	6	5	9	7	1	4	4	3	4	1	1	1	1	1	1	1	1	1	1	1	675

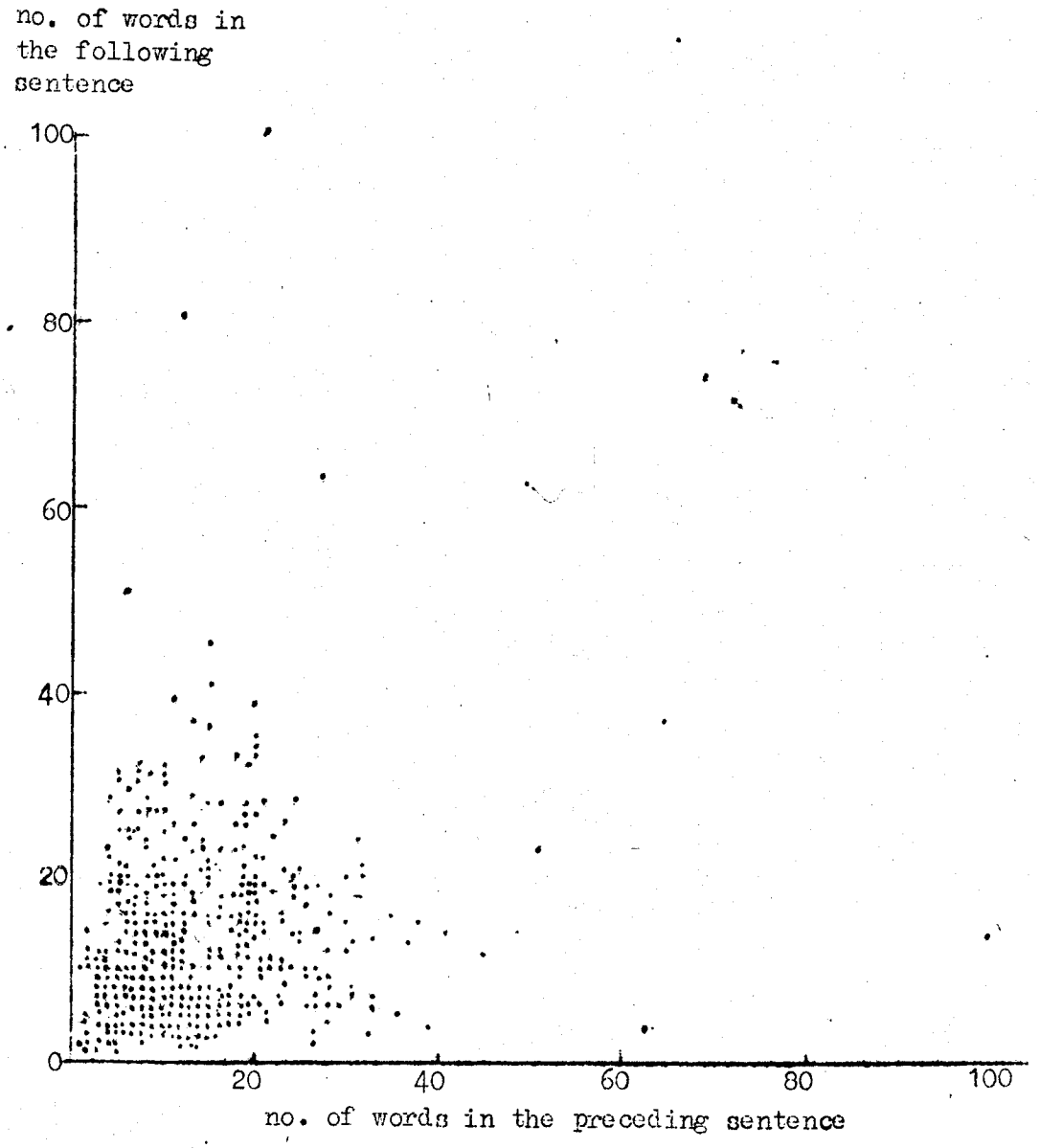


Fig. 9.1(a) : Scatter diagram showing inter-correlations between lengths of two consecutive sentences, based on a complete count of the 675 sentences in Tagore's "Chaturanga", Parts 1 and 2 (Vide Table 9.2).

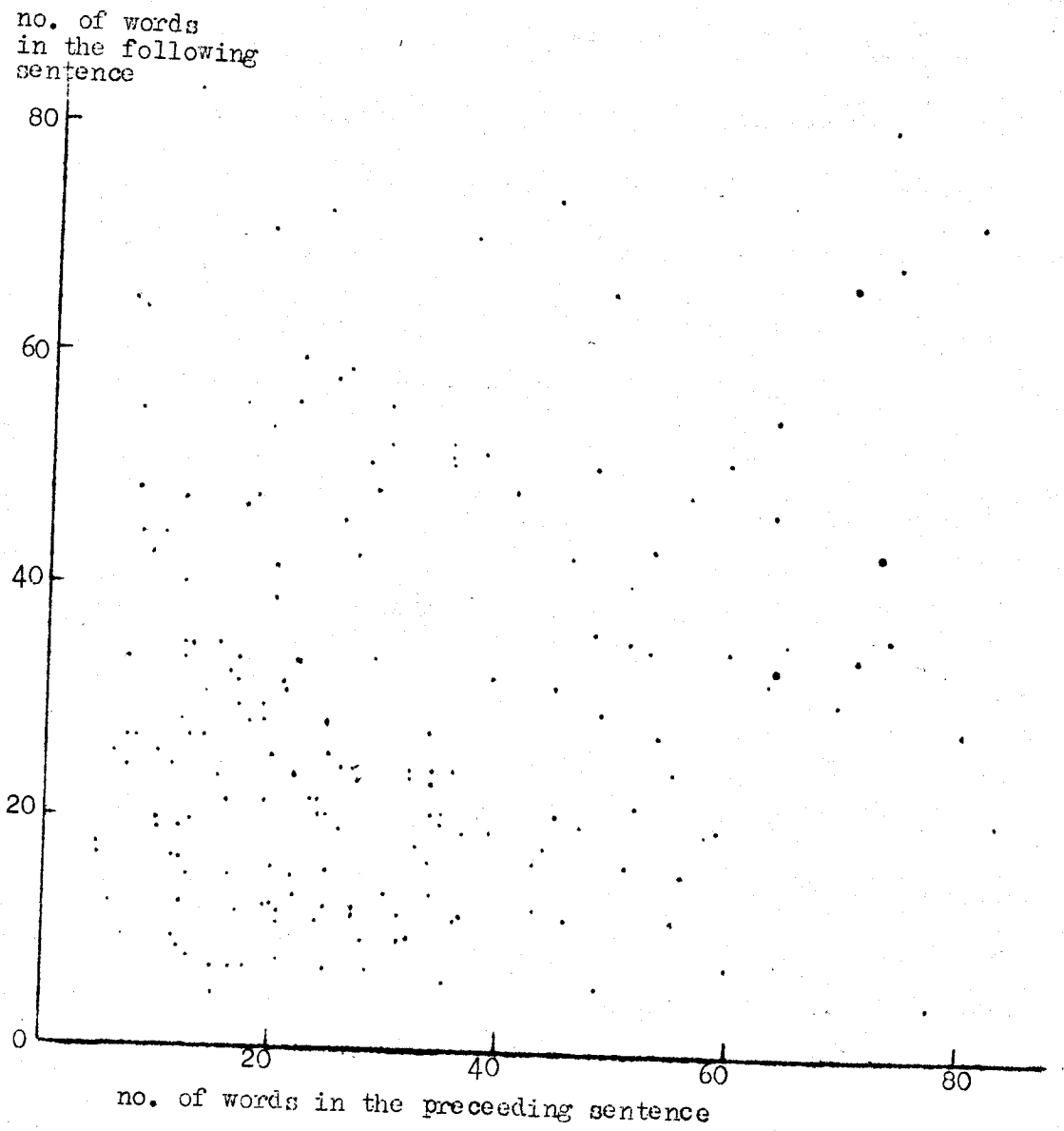


Fig. 9.1(b): Scatter diagram showing inter-correlations between lengths of two consecutive sentences in Extract II from Jane Austen's "Pride and Prejudice" (189 sentences from Chapters 43-44).

9.2.6. Test based on Kendall's rank correlation coefficient (Kendall and Stuart, 1961, pp.478-486) : This is the only test applied which is really sensitive to trend or shifts in the average level. We calculate the rank correlation coefficient $\tau(x_t, t)$ between the values of x_t (here sentence-length) and t, the serial number or the order of occurrence of the sentence in the natural reading order. Actually, if $\tau = \frac{S}{{}^{n}C_{2}}$, the test is applied on the score S, which has properties analogous to those of τ. Under the null hypothesis of randomness, the conditional mean and variance of S are given by :

$$E(S) = 0 \text{ and } V(S) = \frac{n(n-1)(2n+5)}{18}.$$

The distribution tends to normality very rapidly; the approximation is good for $n \geq 10$. Only a correction for continuity has to be made, since S is integral and changes by multiples of 2. If S is too high, a rising trend is indicated, but if S is too low, there is a downward trend.

9.2.7. This test is known to be one of the most powerful nonparametric tests for trend [vide Kendall and Stuart, ibid; Walsh, ibid] though it may not detect oscillations or even trends which are not monotone in character. The slightly more powerful tests given by Walsh (op. cit, p.76) were not applied; the small increase in power does not seem to be worth the trouble.

9.2.8. Since the sentence-length series were rather too long for the straightforward application of this test — the ranking and the calculation of S would have been very laborious — the length of the series was first reduced

by grouping mutually exclusive sets of 3, 5 or 10 consecutive observations[1], say, and calculating the averages or totals for these groups. The test was applied to be series of totals or averages. [vide Walsh, 1962, Chap. 5]. Under the null hypothesis of randomness of the original series, this derived series would also be random. It is hoped that the test remained sufficiently sensitive after the reduction of the length of the series.

9.2.9. Wald-Wolfowitz (1943) test for circular autocorrelation coefficient r_h : The circular autocorrelation coefficient r_h of order h is defined by (h = 1,2, ...,)

$$r_h = \frac{\sum_1^n x_t\, x_{t+h} - (\sum_1^n x_t)^2/n}{\sum_1^n x_t^2 - (\sum_1^n x_t)^2/n} = \frac{R_h - (\sum_1^n x_t)^2/n}{\sum_1^n x_t^2 - (\sum_1^n x_t)^2/n}$$

where $x_{n+k} = x_k$ for k = 1, 2, ..., h. The circular coefficients are used in preference to the noncircular ones, because of simplicity of the sampling theory. If $h \ll n$, and if the series does not have any pronounced trend, the circular coefficients should not give a different picture from that given by the noncircular ones. Also sampling distributions of circular coefficients may be taken as approximations to those for the noncircular coefficients, with, of course, adjustments in the value of n.

1/ If n was not an exact multiple of 3, 5 or 10 (say), a few observations at the end were ignored.

9.2.10. The conditional test for r_h reduces to that for R_h. Under the hypothesis of randomness, the conditional mean and variance of R_1 are given by

$$E(R_1) = \frac{S_1^2 - S_2}{n-1} \text{ and } V(R_1) = \frac{S_2^2 - S_4}{n-1} + \frac{S_1^4 - 4S_1^2S_2 + 4S_1S_3 + S_2^2 - 2S_4}{(n-1)(n-2)} - \frac{(S_1^2 - S_2)^2}{(n-1)^2},$$

where $S_k = \sum_1^n x_t^k$ $(k = 1, 2, 3, 4)$. The conditional distribution of R_1 is also approximately normal for moderate values of n. This provides a test of significance of the coefficient r_1. If h is prime to n, the conditional distribution of R_h and hence the test of significance of r_h is the same as that for R_1 and r_1 respectively. The condition of primeness can be fulfilled by omitting a few observations at either end. But no such omission was done in the present case, when testing the significance of say r_6 with n = 675; the errors are believed to be negligible.

9.2.11. This test also has some optimum properties [Walsh, ibid]. Although it should be capable of detecting both trends and shifts in the average level, as well as oscillations, it is not really efficient for protecting against trends or shifts especially in large samples, but is relatively sensitive to oscillations. The same remark applies to the two Wallis-Moore tests described below.

9.2.12. The circular autocorrelation coefficients $r_1, r_2 \ldots$, are themselves of great interest apart from their use in the above tests.

The expression for $V(R_1)$ gives an expression for $V(r_1)$, and ignoring the condition of primeness, this is also the variance of all the higher order coefficients r_h.

9.2.13. Test based on total number of turning points (p) : Let p be the total number of turning points, that is, relative maxima or minima (peaks or troughs) in the series of length n. For a random series with n terms, the conditional mean and variance of p are given by

$$E(p) = \frac{2}{3}(n-2) \text{ and } V(p) = \frac{16n - 29}{90}$$

Again, the distribution tends to normality fairly rapidly as $n \to \infty$ [Walsh, ibid]. This gives a test of significance for the deviation of observed p from E(p). Generally, p tends to be too low when trend or oscillations are present; but p might conceivably be found to be too high. A continuity correction must, of course, be applied, for p is integral.

9.2.14. For this test as well as for those for r_h described above, the test should strictly be two-sided, but ordinarily the r_h -values tend to be positive and p to fall below E(p). We have applied the two-sided tests in the present case.

9.2.15. χ^2-test based on distribution of phase lengths (Wallis and Moore, 1941) : The phases are the runs up and down, excluding the incomplete run preceding the first turning point and also that following the last turning point. The length of a phase is the number of like

signs (+ or —) in the sequence of first differences of x_t. Let O_1, O_2 and O_3 be the observed numbers of phases of length 1, 2 and 3 or more respectively. Under the hypothesis of randomness, the conditional expectations of O_1, O_2, and O_3 are respectively

$$E(O_1) = e_1 = \frac{5(n-3)}{12}, \quad E(O_2) = e_2 = \frac{11(n-4)}{60} \text{ and } E(O_3) = e_3 = \frac{4n-21}{60}$$

To compare the observed and expected number of phases of different lengths[1/] compute

$$\sum_i (O_i - e_i)^2/e_i$$

and consider it (for large n) as $\frac{7}{6}\chi^2$ (with 2 d.f.) if $\sum(o_i - e_i)^2/e_i \leq 6.3$, and as χ^2 with 2.5 d.f., if otherwise.

9.2.16. It may be noted that the statistic p of para 9.2.13 equals $O_1 + O_2 + O_3 + 1$, and p + 1 is the total number of runs up and down in the whole series.

9.2.17. For the two Wallis-Moore tests as well for the rank-correlation for trend, the problem of ties among observations cropped up here and there, and the randomization method was adopted to solve such problems. The midrank method was not used. All the tests mentioned have the advantage that their application does not have any conditional probability effect on the statistical procedures re-using the same data in any symmetric manner. The randomization method was chosen because it preserves this property exactly (Walsh, ibid).

1/ Here $\sum o_i \neq \sum e_i$ necessarily.

9.3.1. Results of tests of randomness : One may now look at Tables 9.1(a) and (b) which summarise the results of these tests of randomness. The salient features are given below.

9.3.2. Test using $\tau(x_t, t)$: One critical ratio is significant here at the (two-sided) 5% level, viz., that for "Sheser Kavita" extract. Two others are fairly large, about 1.5, viz., those for "Chaturanga", Part 1, and "Bamkimchandra". These results and hence the observed $\hat{\tau}$'s should not be taken too seriously. When the sample size is small, as here, real differences may not come out significant, and unreal differences may appear to be significant. One should note how the critical ratio for "Chaturanga", Parts 1 and 2, combined is much less in absolute value than that for "Chaturanga," Part 1, alone.

9.3.3. Let us take an overall view. Among the 8 independent critical ratios, + and − signs occur 4 times each. Also 7 of these are within the limits -1.96 to +1.96. One might argue that some works or parts of works may have rising trend while others have declining trend, so that the critical ratios would tend to move away from 0 in either direction. Even then one can see that the absolute values of the critical ratios are : 0.169, 0.312, 0.539, 0.631, 0.876, 1.501, 1.559 and 2.040. Sum of squares of the eight critical ratios is 10.427, and this being a χ^2 with 8 d.f.'s is not significant. It does not seem that there is sufficient ground to conclude that some or all the literary works covered here have significant trends.

9.3.4. Wald-Wolfowitz test for circular r_1 : Here one can see that the information from "Chaturanga", Part 1, is being reinforced by that from "Chaturanga", Part 2. If attention is confined to this work alone, the value of r_1 must be considered as significant. The critical ratio for "Sheser Kavita extract" is also appreciable, though not quite significant at the 5% level even using a one-sided test.

9.3.5. The critical ratios for the three Bengali essays are clearly non-significant, remembering the footnote to Table 9.1(b) regarding "Bamkimchandra". The same may be said of the two extracts from "Pride and Prejudice", one giving a significant result, but the other giving a negative critical ratio.

9.3.6. If one ignores the fact that the source texts are very different in nature, one may try to combine the eight independent critical ratios to form an overall conclusion. Since we are interested in a one-sided test, we find the sum of the eight normal deviates, which is 9.056. Under the null hypothesis, this is distributed as $N(0, \sqrt{8})$. So the overall conclusion must be that r_1 is significantly larger than zero. The simple average of the eight r_1's, it may be noted, is of the order of 0.07. This seems to be a reasonable overall estimate, but for individual works the coefficient may be nearer zero or above 0.1.

9.3.7. Wald-Wolfowitz tests for circular r_h (with $h \geq 1$) : If one were to see the figures for "Chaturanga", Parts 1 and 2, by themselves, one would infer that the r_h-values decline after r_2; that r_2 and r_3 are significantly above zero and r_6 nearly so, but r_{10} is almost negligible.

The figures for "Sheser Kavita (extract)" are erratic and partly upset these impressions. The critical ratios are nowhere near significance, show the negative sign, and do not suggest any clear pattern at all. Similar remarks apply to the figures for three Bengali essays and the two extracts from "Pride and Prejudice". On the whole, one should not reach any firm conclusions about the coefficients r_h, save that they are all near zero.

9.3.8. Test based on the total number of turning points (p) : This test gives nonsignificant results in every case. The maximum value of the critical ratio, in the absolute sense, is -1.434 for "Chaturanga", Part 1, and this is not significant even at be one-sided 5% level. All other critical ratios are below 1 in absolute value. Omitting "Chaturanga", Parts 1 and 2, which do not give any independent information, 3 of the 8 critical ratios are negative, and another is exactly zero. It is not necessary to combine the tests in any sophisticated manner. It is very plain that the results of this test are nonsignificant for individual series as well as on the whole, for the critical ratios behave more or less like independent samples from the N(0, 1) population. There is little evidence of any tendency of the observed p-values being smaller (or larger) than E(p).

9.3.9. Wallis-Moore test on the distribution of phase lengths : Although the formal test is an overall χ^2, one may start by comparing the observed and expected frequencies in each of the three length classes (1, 2, 3 or more). One actually finds that the signs of the

deviations are positive and negative more or less equally often in any of cols. (6) to (8) of Table 9.1(a); more generally, large and small deviations have these two signs more or less equally often. So nothing systematic seems to be there even in individual deviations $O_i - e_i$ (i = 1, 2, 3).

9.3.10. As regards the overall χ^2, no χ^2 is significant. The highest value is 4.374 and $\frac{6}{7}$th's of this is 3.75; reading χ^2-tables for 2 d.f., the P-value, i.e., the upper tail probability is about 17%. Not only this. Even without calculating such exact probabilities for each of the eight independent χ^2's, it can be seen that the P-values are fairly well spread over the range 0 to 1. They are rather too small, on the average, for the simple average of the eight χ^2's is 1.65, $\frac{6}{7}$th's of which would be about 1.4, while the average of χ^2 with 2 d.f. is 2. Combination of these χ^2's by standard methods would give a non-significant result.

9.3.11. Fractile Graphs : For the sake of interest we present here some "fractile graphs" [Figs. 9.2(a)— (b)] which enable one to appreciate the randomness of the series in a semi-intuitive manner. The reader will find a discussion of fractile graphs in Mahalanobis (1958, 1960). The relevant figures are shown in Table 9.3.

Table 9.3: Average sentence-length in terms of words by fractile groups based on serial order of sentences, for two selected series of sentence-lengths

fractile group	sub-sample 1		sub-sample 2		combined	
	serial nos. of sentences	average length	serial nos. of sentences	average length	serial nos. of sentences	average length
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(a) Chaturanga : Parts 1 and 2					
1	1- 16	13.12	17- 32	13.75	1- 32	13.44
2	33- 48	10.50	49- 64	13.37	33- 64	11.94
3	65- 80	15.75	81- 96	11.56	65- 96	13.66
4	97-112	17.81	113-128	18.25	97-128	18.03
5	129-144	9.75	145-160	12.44	129-160	11.10
6	161-176	10.00	177-192	12.87	161-192	11.44
7	193-208	13.56	209-224	8.56	193-224	11.06
8	225-240	10.94	241-256	15.44	225-256	13.19
9	257-272	11.87	273-288	12.75	257-288	12.31
10	289-304	10.44	305-320	14.25	289-320	12.34
11	321-336	8.12	337-352	10.00	321-352	9.06
12	353-368	13.12	369-384	12.37	353-384	12.74
13	385-400	5.81	401-416	12.81	385-416	9.31
14	417-432	18.19	433-448	8.87	417-448	13.53
15	449-464	8.00	465-480	11.81	449-480	9.90
16	481-496	15.94	497-512	17.12	481-512	16.53
17	513-528	21.81	529-544	13.19	513-544	17.50
18	545-560	7.50	561-576	13.94	545-576	10.72
19	577-592	13.62	593-608	12.25	577-608	12.94
20	609-624	8.94	625-640	9.62	609-640	9.28
	(b) Pride and Prejudice : Extract II					
1	1- 9	22.89	10- 18	32.44	1- 18	27.67
2	19- 27	38.78	28- 36	26.00	19- 36	32.39
3	37- 45	33.11	46- 54	32.44	37- 54	32.78
4	55- 63	28.44	64- 72	17.44	55- 72	22.94
5	73- 81	25.56	82- 90	20.22	73- 90	22.89
6	91- 99	20.89	100-108	21.00	91-108	20.95
7	109-117	31.33	118-126	36.33	109-126	33.83
8	127-135	26.78	136-144	24.67	127-144	25.73
9	145-153	26.78	154-162	47.00	145-162	36.89
10	163-171	32.78	172-180	31.56	163-180	32.17

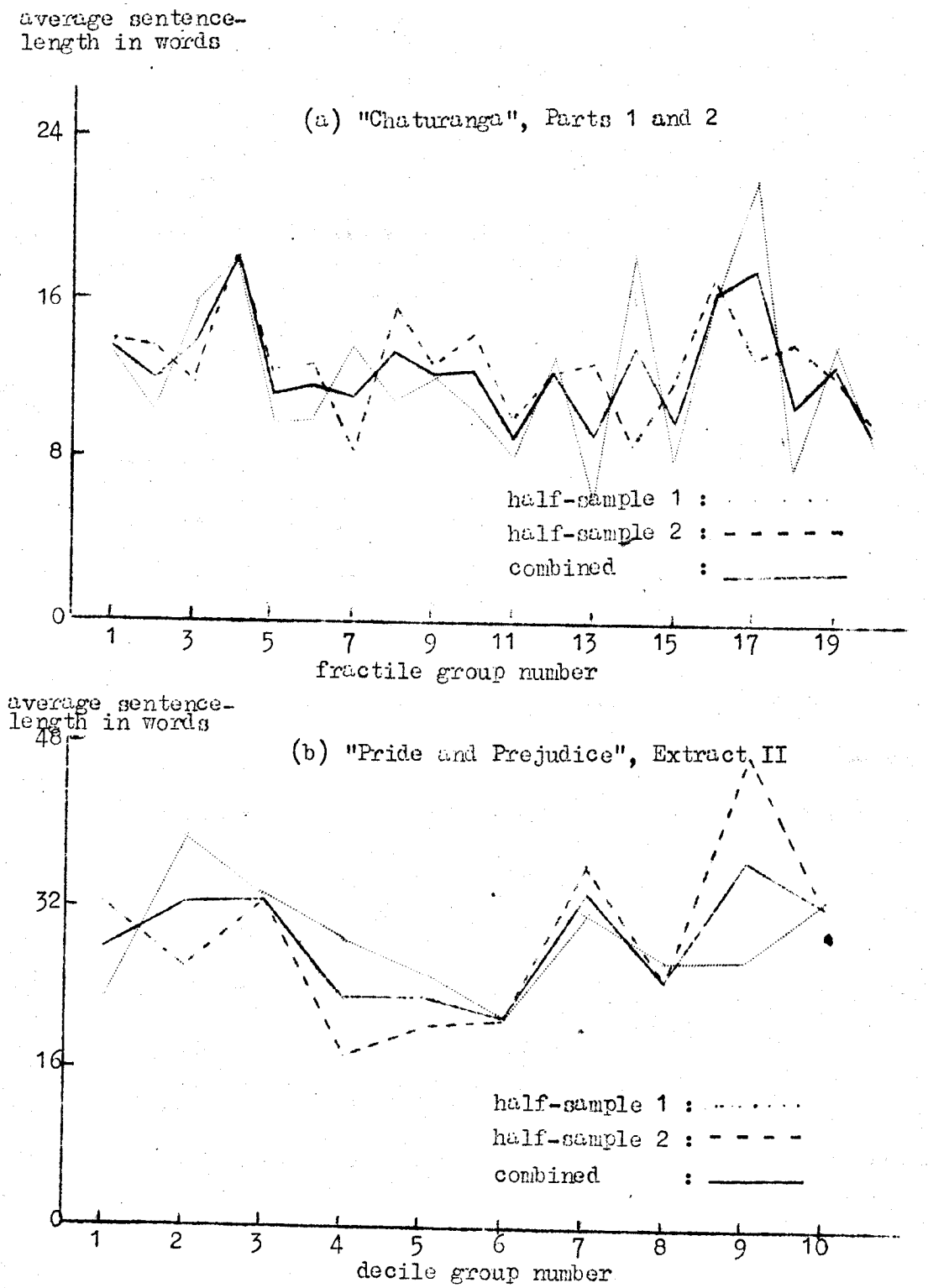


Fig.9.2 : Fractile graphs for sentence-length in terms of words based on a complete count of (a) 675 sentences in "Chaturanga", Parts 1 & 2, and of (b) Extract II from "Pride and Prejudice" (189 sentences), the grouping being based on the serial order of occurrence [Vide Table 9.3].

9.4.1. Autocorrelations for complete works : For estimating autocorrelation coefficients based on complete works, one must either carry out complete counts or depend on representative samples. As regards complete counts, some results are already presented in Sections 9.2 and 9.3, but most of the texts are very short; only the results for "Chaturanga", Parts 1 and 2, have some interest, representing half of a novel, however short. As regards representative samples, the probability samples of sentences could be used, but very little work was done, in view of two types of difficulties. If one used the sentence-pairs observed within the sample sentence-clusters, the probabilities of inclusion would be somewhat different for pairs with both sentences terminating on the same line from the probabilities of pairs where the two sentences terminate on different lines. Ordinary processes of estimation would therefore give biased estimates. Even otherwise, the sampling theory would be complicated, as seen in Section 5.3 of Chapter 5 in connection with corresponding studies on word-length. In the word-length studies, the fairly large sizes of samples enabled us to make fairly reasonable assumptions. In the present case, such assumptions might be hazardous, in view of the small sizes of samples. Table 8.2 shows that for most of the works, the probability samples gave around 200 sample sentences from 60 to 70 sentence-clusters. This would give less than 150 sentence-pairs on which to base the estimation of r_1. Clearly, such estimates would be inconclusive even for the present purposes.

9.4.2. We present, however, Table 9.4, showing the joint distribution of lengths of consecutive sentences in "Krishnakanter Will". This is built up from the sentence-pairs observed on the sentence-clusters comprising the probability sample from this work. It is hardly necessary to show the scatter diagram for this joint distribution. Clearly, there is no marked linear or non-linear correlation between consecutive lengths, as seen earlier from Table 9.2 and Figs. 9.1(a) and (b).

9.4.3. The ordinary process of calculation gives $r_1 = 0.185$ from Table 9.4. The four subsample estimates are : 0.336, — 0.010, 0.098 and 0.294. The combined estimate is nearly significant at the one-sided 5% level, but as the subsample divergence is very large, it is not advisable to say anything about the precise magnitude of the true autocorrelation coefficient.

9.5.1. Systematic samples versus probability samples : As in word-length studies (vide Chapter 2, Section 2.6-7) systematic samples of sentences were drawn from three works in Bengali prose, and compared with probability samples from the same works. It was found that the systematic samples resembled the probability samples, which implies a kind of randomness of the series of sentence-lengths. This experimentation was carried out on a small scale, however, on three works only, as already stated. The works are :"Krishnakanter Will","Yogayog"and "Kavi Shri Ramakrishna" The sentence-length distributions have been presented in Chapter 8 (Table 8.4) and inferences drawn there on the joint evidence of both types of samples.

Table 9.4: Joint distribution of lengths of consecutive sentences based on probability sample from Krishnakanter Will" (no. of randomly selected lines : 100, no. of sentence-clusters : 81)

no. of words in the preceding sentence (x)	no. of words in the following sentence (y)																									total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	22	23	24	26	29	34	
1.)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)
1		1		1	1	1								1												
2	2	1	1		1	2				1					2			1								5
3		3	3	1	2	6				1			1						1	1						11
4	1	3	2	2	1		6	2	1	3				1												19
5	1		2	3	4	2	1	4			2	1		1					1					1		22
6		2	2	2	2		3			1	1	5					1								1	23
7		1	1	1	2	3	1	2		3	1		1		1										1	20
8			2	1	1	2			1		1			2					1	1			1			18
9				1	2	3	1	3					1													12
10			5	1	3	2	1	1		1		1	1		1											11
11				1	1	1	2							1		2										17
12			1				1	1	2		1															8
13						1		2																	1	7
14		1	3				1		1	1								2			1					6
15			2	1		1	1	1				1								1						8
16			1								1															7
17				1	2				1																	2
18				1				2		1					1			1								4
19						1								2												6
20					1					1				1												3
21				1																						3
22									1																	1
23															1											1
24														1												1
26																1										1
27													1													1
34																					1					1
42									1																	1
total	4	12	25	18	23	25	18	18	8	13	7	8	5	10	6	3	1	4	3	3	1	1	1	1	2	220

9.5.2. As in word-length studies, the systematic samples were drawn by following simple numerical rules, which made such sampling much more expedient than drawing the probability samples. One might select, for example, the 4th line from top of every odd-numbered page, and note the lengths of sentences, if any, terminating on the selected lines. Four rules of this type were adopted for sampling from each of the three works. The rules were so framed that when all the sample lines were considered together, the first rule gave those occupying the 1st, 5th, 9th,..., positions in the natural reading order, the second rule those having positions 2, 6, 10 ....., and so on for the remaining subsamples.

9.5.3. The systematic sample of lines (hence sentences) was split up into four independent and interpenetrating subsamples (IPNS), the first subsample (rule I) comprising sample lines occupying the 1st, 5th, 9th, ..., positions in the natural reading order, the second subsample (rule II), those in positions 2, 6, 10, ..., and so on. For fractile graphical analysis and other purposes, the subsamples of probability and/or systematic samples were sometimes merged to get two halfsamples. It has been assumed that the sampling error of a combined systematic sample is indicated by the divergence between its subsamples or half-samples[1/].

1/ Strictly speaking, the concept of sampling error cannot be applied since the systematic sample has no element of probability in it.

9.5.4. The probability samples included clusters of three consecutive lines each with one or more sentences terminating on them; but the systematic samples used only the sample lines and preceding or following lines were not brought into the picture.

9.5.5. That probability and systematic samples from the same work give sensibly the same distribution of sentence-length would be evident from the cumulative distributions presented in Table 9.5. The two-sample Kolmogorov test was applied to compare the two methods of sampling and the results are shown in the bottom rows of the table. In all three cases the test shows close agreement between the two sets of estimates. That the variate is not perfectly continuous or the samples are not unrestricted random, cannot have mattered much in the present situation.

9.5.6. Since the subsamples of the systematic samples were drawn in different ways we also compared each subsample of the systematic sample with the combined probability sample from the same work. The results of the K-test are shown in Table 9.6. The K-test gives non-significant result in every case, and also, it may be safely added, on the whole.

Table 9.5: Cumulative distributions of sentence-length based on probability and systematic samples from three works in Bengali prose.

length (no.of words)	cumulative percentages of sentences					
	"Krishnakanter Will"		"Yogayog"		"Kavi Shri-Ramakrishna"	
	probability sample	systematic sample	probability sample	systematic sample	probability sample	systematic sample
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1.99	1.03	0.65	0.72	0.28	0.61
2	6.64	5.13	2.61	3.96	5.83	3.47
3	17.28	12.56	7.17	9.72	15.56	13.26
4	25.25	23.59	13.03	16.55	28.61	24.90
5	34.55	33.59	21.82	21.95	41.11	40.41
6	44.85	44.62	28.99	29.86	51.39	53.26
7	53.82	51.54	38.44	37.41	62.22	64.29
8	61.13	57.18	46.58	46.76	70.56	71.02
9	65.45	64.10	55.38	52.16	76.94	77.55
10	72.09	71.03	60.26	57.92	81.67	83.67
11	75.42	76.92	65.47	65.11	86.39	87.55
12	78.74	81.54	71.01	70.87	89.44	91.02
13	81.73	83.85	77.52	75.19	90.83	92.86
14	85.38	86.92	80.78	77.71	92.22	95.10
15	88.70	88.72	84.04	82.03	95.83	95.92
16	89.70	90.00	87.95	85.99	96.94	97.14
17	91.03	91.28	89.90	88.51	97.22	97.76
18	93.02	92.05	91.53	90.67	97.78	98.16
19	94.68	92.56	93.16	92.11	98.06	98.37
20	95.68	93.33	95.11	93.19	98.06	98.98
21	96.01	94.36	96.09	93.91	98.33	99.18
22	97.01	94.62	96.74	94.63	98.89	99.39
23	97.67	95.13	97.39	95.35	99.17	99.39
24	98.01	95.64	98.05		99.44	99.59
25		95.90		95.71	99.72	99.59
26	98.34	96.15	98.70	96.43		
27	98.67	96.92		96.79		99.80
28		97.18	99.02	97.87		
29	99.00	97.44		98.23		100.00
30				98.59		
31		97.69	99.35		100.00	
32				98.95		
33		97.95		99.67		
34	99.67					
37				100.00		
38		98.20				
39		98.46	99.67			
40		98.72				
41		99.23				
42	100.00					
43		99.49				
50			100.00			
51		99.74				
72		100.00				
no.of sample sentences	301	390	307	278	360	490
K-distance:	4.71%		3.52%		3.71%	
-do- P-value	>> 20%		>> 20%		>> 20%	

Table 9.6: Kolmogorov test for comparing sentence-length distributions from the combined probability sample and from individual subsamples of the systematic sample, separately for three works in Bengali prose*

work	no. of sentences in combined sample		K-distance (%) between combined probability sample and			
	prob.	syst.	sub-sample 1 of syst. sample	sub-sample 2 of syst. sample	sub-sample 3 of syst. sample	sub-sample 4 of syst. sample
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. "Krishnakanter Will"	301	390	8.28	13.35 (15%)	9.11	7.75
2. "Yogayog"	307	278	7.26	7.22	5.34	9.67
3. "Kavi Shri-Ramakrishna"	360	490	5.78	2.68	6.59	8.61

* The P-value is shown inside brackets in the one case where P is not much greater than 20%.

9.5.7. The same result can be seen in another way. Table 9.7 shows the decile group averages of sentence-length, separately for the three works and for the two types of samples. In each case, estimates are given for two halfsamples and for the combined sample. These averages are plotted in the form of fractile graphs in Figs. 9.3(a), (b) and (c). [Vide Mahalanobis, 1958, 1960 for a general account of fractile graphical analysis.] In each case, the graphs for the two types of samples do not show much significant "separation" : The error areas for the two types of samples overlap to a considerable extent. The only divergence worth noting is that for "Krishnakanter Will", the partial separation seems to be nearly significant at the extremes of the range.

Table 9.7: Decile group averages of sentence-length, separately by half-samples and combined, for probability and systematic samples from three works in Bengali prose

decile group (per cent)	average sentence-length in words: probability sample h.s.1	probability sample h.s.2	probability sample comb.	systematic sample h.s.1	systematic sample h.s.2	systematic sample comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
			"Krishnakanter Will"			
0 - 10	2.08	2.19	2.14	2.45	2.33	2.38
10 - 20	3.16	3.39	3.27	3.90	3.61	3.74
20 - 30	4.03	4.92	4.47	4.64	4.64	4.64
30 - 40	5.18	5.91	5.54	5.76	5.54	5.64
40 - 50	6.32	6.70	6.51	6.39	6.67	6.54
50 - 60	7.54	7.72	7.62	8.12	8.13	8.13
60 - 70	8.97	9.60	9.45	9.30	9.84	9.59
70 - 80	10.87	11.98	11.37	10.89	11.49	11.22
80 - 90	14.38	14.43	14.44	13.29	14.52	13.92
90 -100	21.87	22.29	22.06	26.44	25.97	27.14
0 -100	8.44	8.92	8.68	9.12	9.27	9.29
no. of sample sentences	152	149	301	182	208	390
			"Yogayog"			
0 - 10	2.60	2.85	2.56	3.32	2.55	2.96
10 - 20	3.72	4.92	4.34	4.89	4.56	4.70
20 - 30	5.48	6.15	5.82	5.59	6.64	5.92
30 - 40	6.91	7.54	7.26	6.79	7.41	7.16
40 - 50	8.00	8.69	8.32	7.99	8.57	8.34
50 - 60	9.72	10.31	9.99	9.52	9.48	9.46
60 - 70	11.31	11.85	11.49	11.58	11.17	11.43
70 - 80	12.96	14.23	13.62	13.25	13.04	13.15
80 - 90	15.99	16.93	16.35	15.68	15.83	15.73
90 -100	23.68	25.39	25.22	21.71	23.99	22.82
0 -100	10.04	10.89	10.50	10.03	10.30	10.17
no. of sample sentences	162	145	307	148	130	278
			"Kavi Shri-Ramakrishna"			
0 - 10	2.26	2.49	2.39	2.56	2.62	2.59
10 - 20	3.24	3.65	3.44	3.57	3.79	3.67
20 - 30	4.00	4.42	4.14	4.29	4.74	4.51
30 - 40	4.87	5.12	5.00	5.00	5.11	5.00
40 - 50	5.66	6.00	5.89	5.82	6.00	5.96
50 - 60	7.00	6.72	6.86	6.66	6.69	6.67
60 - 70	7.98	7.59	7.78	7.55	7.62	7.57
70 - 80	9.42	9.05	9.25	8.94	9.34	9.13
80 - 90	11.71	10.98	11.26	10.79	10.97	10.87
90 -100	18.02	15.53	16.91	16.08	15.61	15.81
0 -100	7.42	7.16	7.29	7.13	7.25	7.18
no. of sample sentences	182	178	360	251	239	490

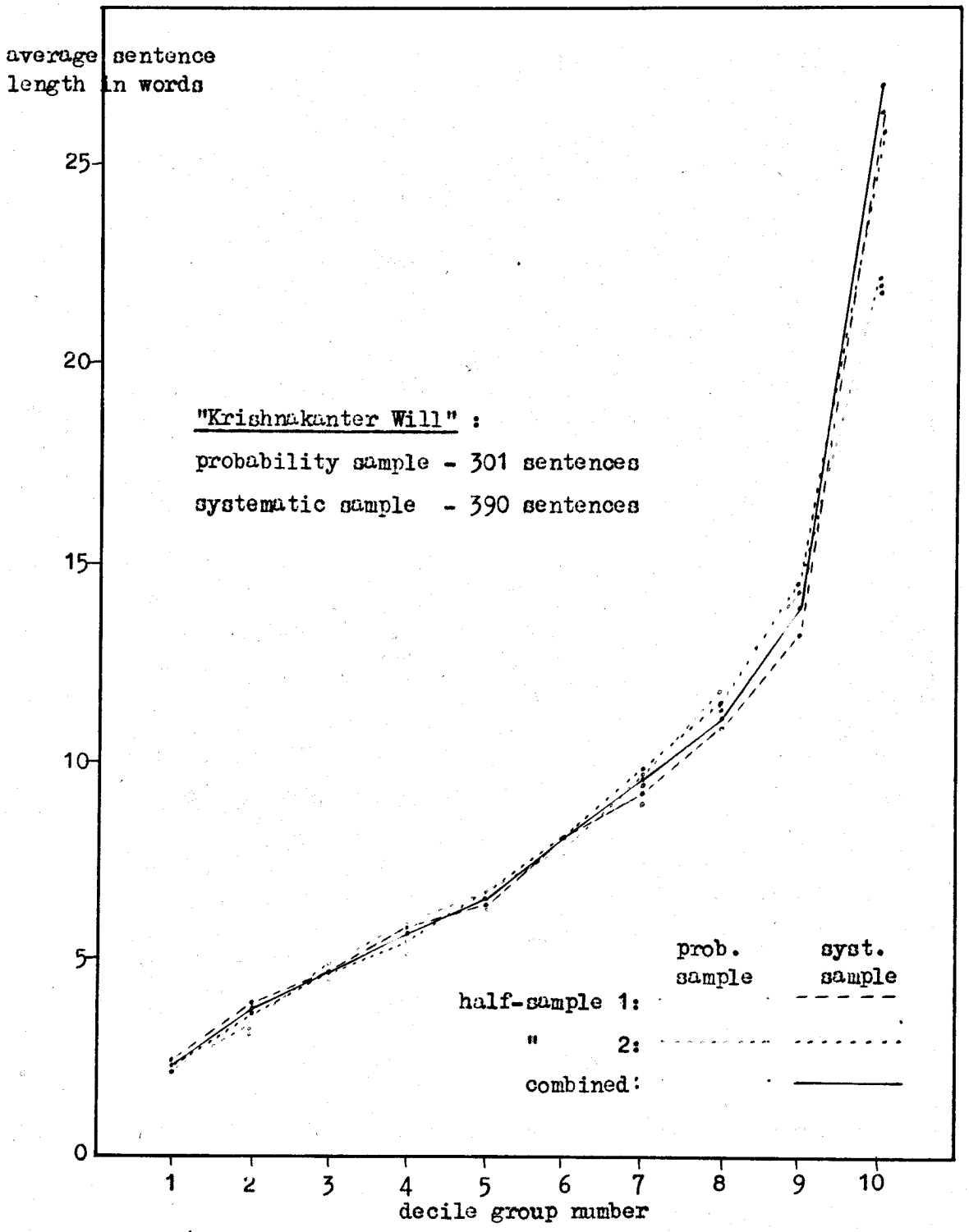


Fig.9.3(a): Fractile graphs for sentence-length in terms of words showing agreement between probability and systematic samples from "Krishnakanter Will" [Vide Table 9.7].

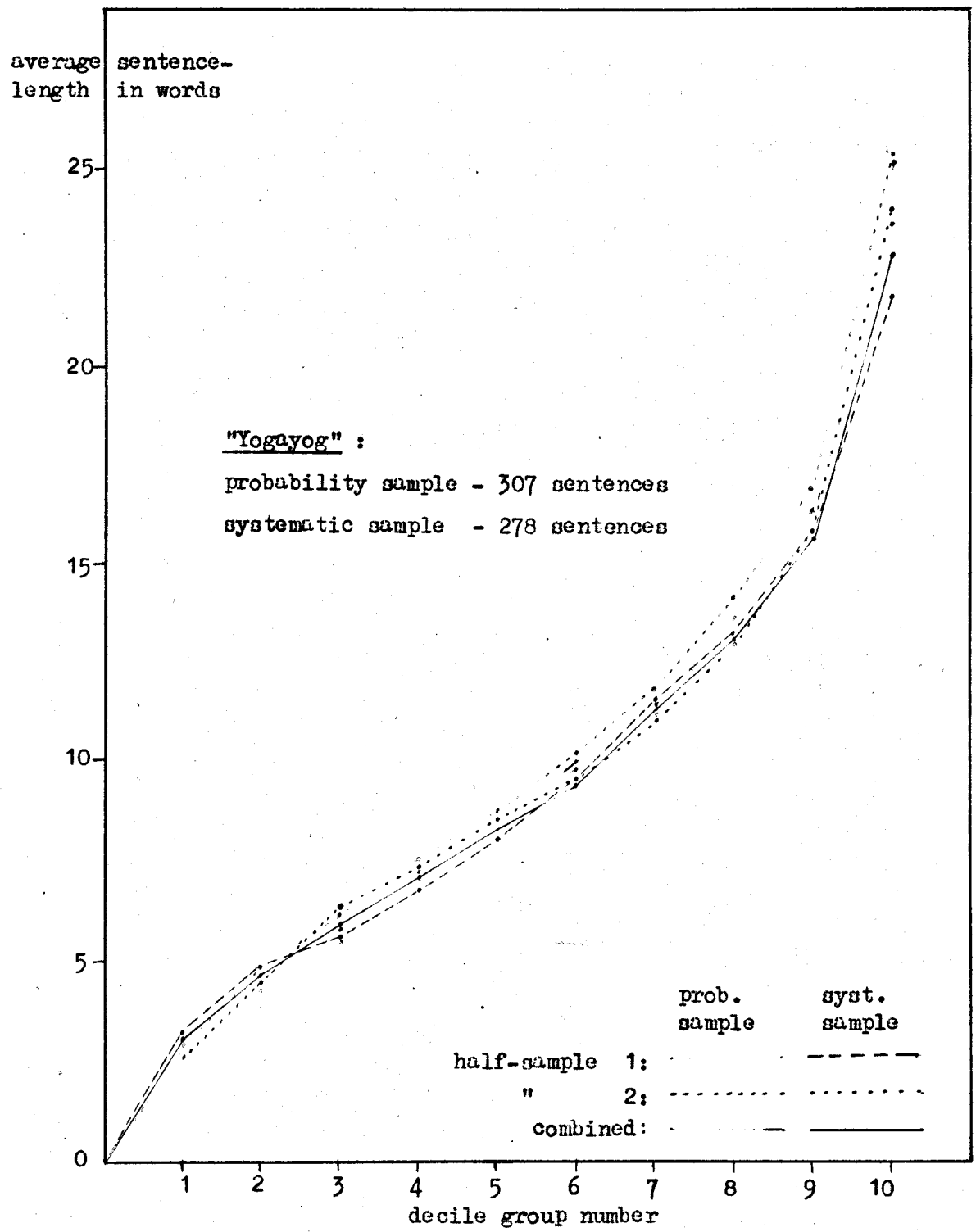


Fig.9.3(b): Fractile graphs for sentence-length in terms of words showing agreement between probability and systematic samples from "Yogayog" [Vide Table 9.7].

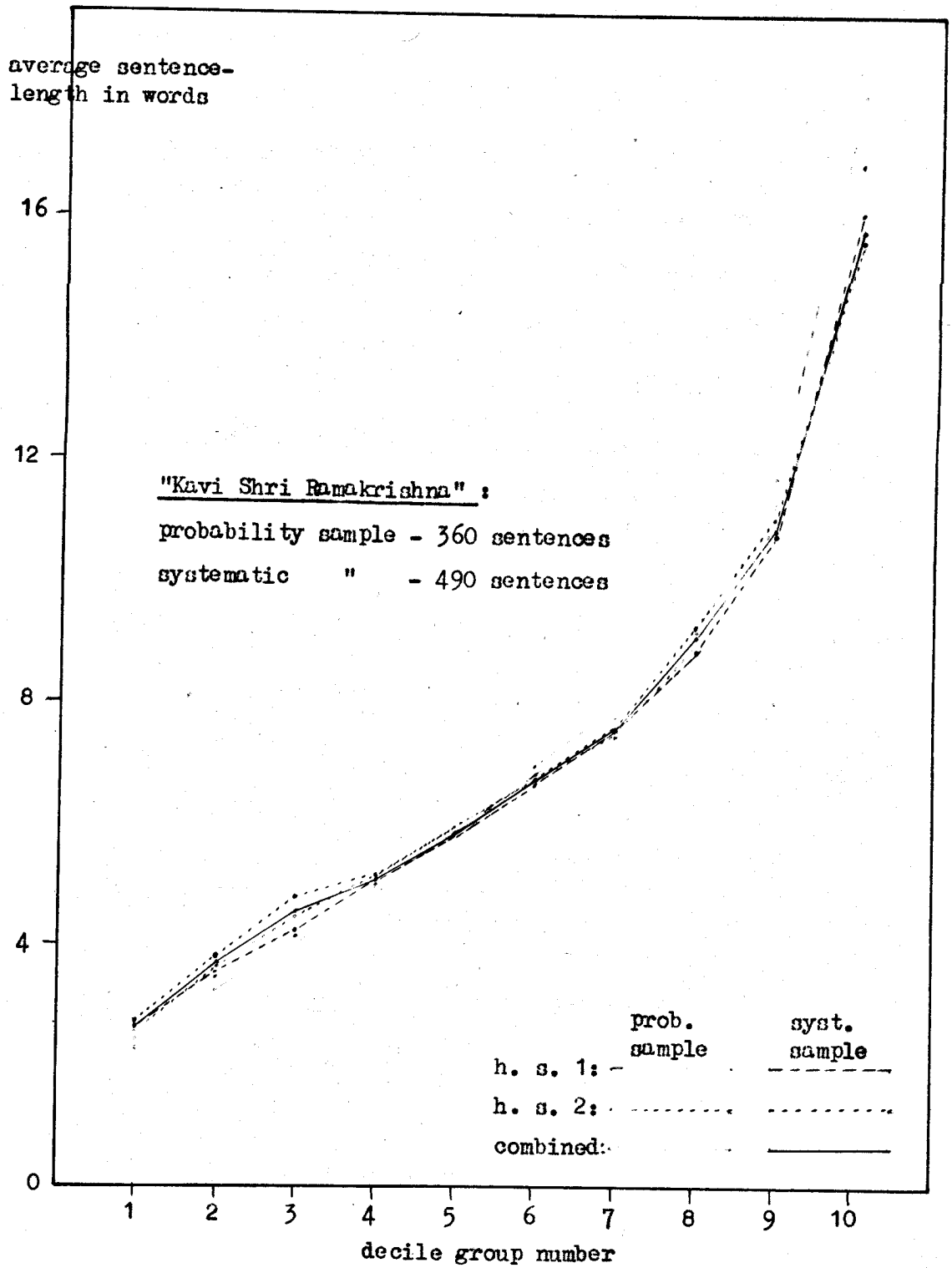


Fig. 9.3(c): Fractile graphs for sentence-length in terms of words showing agreement between probability and systematic samples from "Kavi Shri Ramakrishna" [Vide Table 9.7].

9.5.8. The fractile graphs also indicate that the sampling errors of the systematic samples are of the same order of magnitude as those of probability samples of the same size. Such fine points cannot be decided on the small body of evidence examined here. For the sake of interest, we present in Table 9.8 the Kolmogorov distances between each pair of subsamples, separately for the probability and the systematic samples, from each of the three selected works. The table seems to support the broad conclusion stated in the opening sentence of this paragraph.

Table 9.8: Kolmogorov distances between sentence-length distributions estimated from different pairs of subsamples, separately for probability and systematic samples from three works in Bengali prose*

subsamples compared	probability sample			systematic sample		
	"Krishna-kanter Will"	"Yogayog"	"Kavi Shri Rama-krishna"	"Krishna-kanter Will"	"Yogayog"	"Kavi Shri Rama-krishna"
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 vs 2	13.82	15.46	10.95	17.44 (10%)	10.79	7.61
1 vs 3	6.22	10.91	11.33	15.66 (18%)	11.52	8.98
1 vs 4	8.94	14.81	18.22 (10%)	8.93	13.89	11.85
2 vs 3	16.16	10.07	8.63	11.64	11.48	5.77
2 vs 4	14.81	9.17	11.89	12.58	11.74	7.94
3 vs 4	8.64	8.11	17.65 (12½%)	15.73 (18%)	13.34	11.21
average	11.43	11.42	13.11	13.66	12.13	8.89
no. of sentences (combined sample)	301	307	360	390	278	490

* The P-values are given inside brackets in cases where they are not greater than 20%.

9.6.1. Some evidence on homogeneity of works : Some light on the nearly random nature of the sentence-length series is thrown by certain material already presented in Chapter 8. Actually, these point to the homogeneity of certain works in respect of the frequency distribution of sentence-length.

9.6.2. Consider, first, the sentence-length distributions for Parts 1 and 2 of "Chaturanga", presented in Table 8.5 in Chapter 8. As noted earlier, these are obtained by complete counts. If we apply the two-sample Kolmogorow test for judging the homogeneity of the two distributions, we get K = 8.73% [at x (sentence-length) = 9]. Since n_1 = 380 and n_2 = 295, this value is not significant; the P-value is approximately 16%. But the K-test should be conservative in the present situation; sentence-length is a discrete variate and the two "samples" of sentences are not perfectly random in view of the positive autocorrelations between successive sentence-lengths [vide Table 9.1(b)].

9.6.3. We compare next the sentence-length distribution for the 155-sentence extract from "Sheser Kavita" (Table 8.5, Chapter 8) with that shown in Table 8.3 based on a probability sample of 192 sentences from the whole of the same work. The extract is the chapter entitled "Lavanya Puravritta", and is rather unusual in the work, having very little of conversations. We might apply the two-sample K-test as a very rough criterion, ignoring the points mentioned in the foregoing

paragraph, and also the overlap between the two sets of sentences. The K-distance is found to be 11.95%, at x = 8 ; this is also not statistically significant, the P-value being nearly 18%.

9.6.4. Finally, we compare the two haphazardly selected extracts from "Pride and Prejudice" with 106 and 189 sentences respectively. The sentence-length distributions are presented in Table 8.5 of Chapter 8. The two-sample K-distance is 8.17% at x = 11. This is far from statistically significant; the P-value is far above 20%. For remarks on the applicability of the K-test vide para 9.6.2 above.

9.6.5. It appears that significant differences can only be found by dividing works of the usual size into a small number of fairly long parts. Otherwise, the sample sizes would be too small to show even appreciable differences as significant. But such lengthy parts would tend to be very similar to one another, because different types of patches would tend to be evenly distributed among the small number of lengthy parts. Whereas short patches of shortish or longish words can be statistically detected (vide Chapter 5), short patches of shortish or longish sentences cannot perhaps be detected with equal ease.

Chapter 10 : Correlation between word-length and word-frequency in written English

10.1.1. Introduction : Suppose one carries out a complete count of all words in a sufficiently long text, and prepares a two-way frequency table showing the number f_{xy} of words which are of length x (in terms of the number of syllables or letters or phonemes comprising the word) and which occur y times in the text (x, y = 1, 2, 3, ...). One will find a negative correlation between the two variates, x and y, that is, between the length of a word and its frequency in actual use : The more frequent words are, on the average, shorter, and vice versa. In statistical language, the regressions of y on x and of x on y have negative slopes. There is, however, a good deal of spread around these regression curves, the correlation being far from perfect.

10.1.2. In a general way, this phenomenon is well-known to linguists and is largely ascribed to the forces of linguistic evolution. All words tend to be abbreviated in size, i.e., grounded smaller, in course of time (Jespersen, 1922, Chaps. XVII-XVIII)This force operates more on the high frequency words. As extreme examples of this, we have the so-called truncations, (e.g., 'omnibus' becomes 'bus'). Usually word-frequency is the determinant of word-length, but strictly speaking, the two variates should be regarded as interdependent; for sometimes, as in substitutions (e.g., 'automobile' becomes 'car'), the length of the word influences its frequency (vide, Zipf, 1949,pp.65-66).

10.1.3. A systematic statistical study was carried out by Zipf. His findings are summarised in Zipf (1949). Zipf emphasised the economy of having a negative correlation between word-length and word-frequency[1]. His treatise is a broad exposition of the principle of least effort as the principle governing all spheres of human activities, and this negative correlation served as an illustration of the working of the principle. Zipf presents numerical tables (Zipf, 1949, p.64) for American newspaper English (the Eldridge count) and for Plautine Latin. These tables give the average lengths of words occurring 1,2,3,..... times in the count, and indicate the negative slope of the regression of word-length on word-frequency. According to Zipf, this correlation has been observed in a variety of languages, and "there is little incentive to pursue the matter further quantitatively" (Zipf, 1949, p.65).

10.1.4. Herdan (1956) discusses this correlation at several places of his work. He presents some tables (pp. 74-75, pp. 140-143) showing (implicitly) the regression of word-frequency on word-length for English, German and Russian; the negative correlation is apparent from these tables also. Herdan (1958a) used the form $\bar{y}_x = ax^{-b}$ for the regression of y on x : for some material on conversational English considered by Herdan, with x as the number of letters or phonemes, b was found to be nearly 2.4.

10.1.5. Miller and Newman (1958) studied this correlation in written English in an attempt to explain the rank-frequency relation observed

---

1/ This must have been obvious before Zipf's work.

by Zipf and others. Miller, Newman and Friedman (1958) considered (i) content words, covering nouns, verbs, adjectives plus most of the adverbs, and (ii) function words, covering the rest. They observed that the length-frequency correlation is pronounced in the case of function words, while the frequencies of content words are "relatively independent of length".

10.1.6. All these studies were concerned with the regression of y on x or of x on y. Although highly informative, the regressions do not give a complete picture, and statistical measures seem to be necessary for the strength of the correlation.

10.1.7. The object of this chapter is mainly to study the correlation coefficients (r) and the correlation ratios (η) in some word-counts on (written) English prose. Section 10.2 discusses a matching model and shows that even if the regressions are curvilinear, the coefficient r happens to be much more meaningful in the present context : it indicates the economic effect of the negative correlation between word-length and word-frequency. The usefulness of root word-counts is explained in this connection. Section 10.3 presents some empirical values of r, after describing the word-count material analysed for this study. The effect of r on text-length is demonstrated by considering the three hypothetical cases where the matching of concepts (occurring with observed frequencies) and words (the given letter-combinations) is (i) most economic, (ii) random and (iii) most uneconomic. For all 'particular' words in written English prose, the

minimum attainable value of r — that is, the value in case (i) — is only about -0.20 and not -1, and the text-length is then about 46% of that 'expected' for random matching; the corresponding maxima for case (iii) are + 0.28 and 176% approximately. The actual value of r is -0.14 and the actual text-length is 62% of what could be expected under random matching. Such calculations have been done for all root words, all particular words and also for certain subclasses of these. Section 10.4 presents the correlation ratios η . Section 10.5 considers the problem of estimating r, η etc., in infinitely large word-counts, which could not be solved satisfactorily. The main findings about r, η etc., are also discussed in this Section. Section 10.6 contains some remarks on the form of the regression of y on x and also on the conditional distribution of y given x. Section 10.7 points out the limitations of the present study besides making some concluding observations.

10.2.1. Special Significance of the Correlation Coefficient: Although the two regressions are appeciably curved [vide Section 10.6 infra] the correlation coefficient r seems to be more meaningful than the correlation ratio η ; for r and not η indicates the economic effect on text-length of the negative correlation between word-length and word-frequency.

10.2.2. To see this, consider language as a system of codes used for human communication. Suppose that one is given a set of n words i.e., combinations of letters (say), which are devoid of meanings and

with which one has to match n definite meanings or concepts. Suppose further that in a given text, these different meanings or concepts occur with definite frequencies $y_1, y_2, \ldots\ldots, y_n$, which cannot be altered, just as the number of letters comprising each given word cannot be changed. If the matching of the n meanings and the n words be such that f_{xy} x-lettered words each occur y times in the text, then the total length of the text is $\sum_{x,y} f_{xy} xy$ in terms of the number of letters. If spaces between words are counted as one letter each, the length will be $\sum_{x,y} f_{xy} xy + (n-1)$. It thus appears that if one is interested in the total length of the text, the matching of n meanings and n words is equivalent to filling in the joint distribution of x and y when the marginal distributions of x and y are specified.

10.2.3. Now the correlation coefficient between x and y is given by

$$r = \frac{\frac{1}{n}\sum_{x,y} f_{xy} xy - \bar{x}\,\bar{y}}{\sigma_x\,\sigma_y} \qquad (1)$$

which can be rewritten as

$$\sum_{x,y} f_{xy} xy = n(\bar{x}\,\bar{y} + r\,\sigma_x\,\sigma_y) \qquad (2)$$

This relation shows that given the marginal distributions of x and y the length of the text increases linearly with r. If r = 0, which represents the average for <u>random matching</u> of words and meanings,

the length reduces to $n \bar{x} \bar{y}$. It follows that the ratio

$$\frac{\sum_{x,y} f_{xy} xy}{n \bar{x} \bar{y}} = 1 + r \frac{\sigma_x}{\bar{x}} \frac{\sigma_y}{\bar{y}} = 1 + r \text{ C.V.}(x) \text{ C.V.}(y) \quad . (3)$$

gives the length of the actual text as a fraction of the average length expected for random matching.

10.2.4. The matching model described above is obviously an over-simplification of reality. The major criticism that can be levelled against it is that in most languages, including English, related concepts are conveyed by words having close similarity in structure, which are regarded as particular or inflected forms of the same 'root' word (for instance, 'go', 'goes', 'gone', 'going', are particular forms of the same root 'go'). Our model has ignored this useful feature of most languages, so that related concepts may be matched with words having no structural similarity.

10.2.5. It is possible to meet this criticism by formulating an alternative model which is unfortunately somewhat imprecise. This model is concerned with root words instead of particular words, and considers the matching of root words with root concepts. The text-length is counted after substituting for every individual word in the text the corresponding root word. The affixes ('s', 'ing', 'er' etc.) are thereby ignored. The text thus modified may be called the "reduced" text. The affixes are restored after the matching of words and concepts has been decided. This matching model seems to be fairly realistic,

since related concepts are matched with similar words; but it is concerned with the length of the 'reduced' text[1].

10.2.6. This means that root word counts can be more meaningfully analysed from the present viewpoint than counts on particular words. Some root word counts were therefore utilised in the present study besides others concerned with particular words [vide Section 10.3].

10.2.7. It may be recalled in this connection that from the point of view of linguistics, the English language is comparatively free of inflections. The distinction between root words and particular words may matter more for other languages than in the present investigation on written English.

10.2.8. The value of r for all particular words in written English has been found to be nearly -0.14. Although statistically significant, this value might be regarded as rather small. This impression is not quite correct. The value -0.14 should not be judged against -1, which is the theoretical lower limit of r in ordinary cases; the minimum value actually atainable in the present case is much higher, only about -0.2. Given the present type of marginal distributions of x and y — these are necessarily integers and have very different distributions — the values $r = \pm 1$ cannot be attained by any joint distribution $\{f_{xy}\}$ $(x, y = 1,2, \ldots)$. Consequently, the maximum

---

1/ If frequency counts of affixes were available we might include the affixes in the list of roots and carry out the correlational analysis of root-length and root-frequency. Even then our model would remain imprecise; for one thing, particular forms are not always derived by addition of affixes.

or minimum attainable text-lengths cannot be obtained by putting $r = \pm 1$ in eqn. (2) of para 10.2.3.

10.2.9. Given the marginal distributions of x and y, one can fill in the cells of the two-way frequency table in such a manner that the word ranked j in ascending order of length is matched with the meaning ranked j in descending (ascending) order of frequency (j= 1,2,....,n). The hypothetical distribution would represent the most economic (uneconomic) matching of words and meanings. The text-length $\sum_{x,y} f_{xy}\, xy$ calculated from this distribution would give the minimum (maximum) attainable text-length and hence the minimum (maximum) attainable value of r.

10.3.1. Empirical values of the correlation coefficient : We shall first describe the word counts on written English prose which have been utilised in the present analysis. Table 10.1 gives some general information about the material.

10.3.2. We first describe the Miller, Newman and Friedman count (1958) of 36299 words (5537 distinct words). The results of this count are presented in the convenient form of a two-way distribution of word-length in letters (x) and word-frequency (y) (x,y = 1,2,3,....,). Actually, two such distributions are given, one for 5180 content words occurring 14877 times and the other for 357 function words accounting for 21422 word-tokens (vide para 10.1.5). The count was based on three widely different types of text material : (i) The King James Version of the Bible, (ii) 'Talks to Teachers' by William James, and (iii) the department entitled, 'The Atlantic Reports' of the 'Atlantic Monthly' The great advantage of this count is that the complete material is available upto words occurring even once. (Cf. the Dewey count described in the following para). This rare virtue is possessed by the Eldridge count also (vide Dewey, 1923, pp. 3-4 of 1950 edition) but this latter was not covered, being of about the same size as the Miller et al count.

Table 10.1 : General information about the word-counts analysed

word-count material			number of		mean of		s.d. of		c.v. of	
author and source	texts sampled	class of words	word types (n)	word-tokens ($n\bar{y}$)	word length in letters ($\bar{x}$)	word frequency ($\bar{y}$)	word length in letters (σ_x)	word frequency (σ_y)	word length in letters (c_x)	word frequency (c_y)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Miller et al (1958)										
(a) Table II	Bible, 'Talks to Teachers', 'Atlantic Monthly'	content words	5180	14877	7.326	2.872	2.430	5.050	0.332	1.758
(b) Table I	-do-	function words	357	21422	5.392	60.006	2.202	196.485	0.408	3.274
(c) Table I,II	-do-	all words	5537	36299	7.201	6.556	2.462	52.057	0.342	7.940
2. Dewey (1923) :										
(a) Table 3	text representing modern English prose	particular words with $y \geq 10$	1026	78634	5.552	76.641	2.012	329.029	0.362	4.293
(b) Table 4	-do-	root words with $y \geq 10$	1132	87380	5.384	77.191	1.782	315.436	0.331	4.086
3. Yule (1944) :										
(a) Appendix I	Pilgrim's Progress, Pt.I	nouns	1005	4047	6.389	4.027	2.262	9.843	0.354	2.444
(b) Appendix II	-do- Pt.II	-do-	1020	4016	6.259	3.937	2.204	8.817	0.352	2.240
(c) Appendix III	Mr. Badman	-do-	1030	3992	6.662	3.876	2.460	10.660	0.369	2.750
(d) Appendix IV	Holy War	-do-	996	4001	6.591	4.017	2.294	11.369	0.348	2.830
(e) Appendix I-IV comb.	above-mentioned four works of Bunyan	-do-	2249	16063	6.770	7.142	2.398	23.340	0.354	3.268

10.3.3. The second count analysed was that due to Dewey (1923) based on 100,000 words. This very well-known count was based on carefully selected and diversified material representing modern English prose (written). Two frequency lists given by Dewey were analysed : first, the list of 1026 particular words occurring more than 10 times in the count, covering in all 78,634 word occurrences; and second, the list of 1132 root words occurring more than 10 times, and accounting for 87380 word-tokens among the 100,000[1/].

10.3.4. For the sake of interest, we also analysed Yule's count (1944) on nouns in four works by John Bunyan. These counts are available as four appendices to Yule's work. The works concerned are "Pilgrim's Progress, Part I and Part II" (separately), "Mr. Badman" and "Holy War". About 4000 nouns were counted in each work giving roughly 1000 distinct nouns in each case. (Table 10.2 presents the joint distribution of x and y based on all the four counts taken together.) Here also the material is complete in the sense that nouns occurring even once in the count are shown in the list.

10.3.5. Yule's count on nouns was based, more or less, on the root word concept [vide Yule, 1944, pp. 32-33]. Miller et al followed the more common procedure and kept every variant separate.

1/ There are several small discrepancies between Dewey's Tables4 and 6, both relating to root words. Table 4 was used in such cases.

Table 10.2 : Distribution of nouns by length in letters (x) and total number of occurrences (y) in Yule's sample counts on four works of Bunyan (viz., 'Pilgrim's Progress, Parts I and II', 'Mr. Badman' and 'Holy War') taken together.

no. of occurrences (y)	number of letters (x)														total	mean length ($\bar{x}_y$)
	2	3	4	5	6	7	8	9	10	11	12	13	14	>14		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1	-	34	114	113	135	148	114	98	64	67	26	10	9	1[a]	933	7.208
2	-	19	50	47	54	56	42	38	34	17	10	3	1	1[b]	372	6.992
3	1	6	36	31	28	29	25	17	9	2	6	2	-	-	192	6.526
4	-	8	23	18	20	20	15	10	10	2	-	4	1	-	131	6.595
5	-	2	12	17	19	18	9	5	4	4	1	-			91	6.505
6	-	-	15	16	12	14	6	5	4	4	-	1			77	6.494
7	-	-	4	9	10	5	9	4	5	1	-	1			48	7.042
8	-	4	7	8	9	6	4	1	2	-	-	-			41	5.780
9	-	2	9	5	4	3	3	4	1	2	1	-			34	6.353
10	-	1	3	7	6	3	1	4	-	-	-	1			26	6.307
11	-	1	2	5	7	4	5	2	-	2					28	6.643
12	-	2	5	2	2	1	4	2	-	2					20	6.350
13	-	2	5	3	4	5	2	2	-						23	5.826
14	-	1	5	2	2	2	-	-	1						13	5.308
15	-	-	2	4	3	3	1	1	1						15	6.267
16	-	1	3	6	-	1	4	-	-						15	5.600
17	-	1	4	-	2	1	1	-	-						9	5.111
18	-	2	3	1	2	1	3	-	1						13	5.846
19	-	-	3	2	-	-	-	1							6	5.167
20	-	-	1	4	2	1	1	-							9	5.667

(contd.)

Table 10.2 : (Contd.)

no. of occurrences (y)	number of letters (x)														mean length ($\bar{x}_y$)
	2	3	4	5	6	7	8	9	10	11	12	13	14	>14 total	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
21	-	3	2	2	3	2	2	2							16	5.812
22	-	-	2	3	-	1	-	-							6	5.000
23	-	-	2	1	-	1	1	-							5	5.600
24	-	-	1	2	1	-	-	1							5	5.800
25	-	-	-	2	1	1	-	-							4	5.750
26	-	-	3	3	2	1	-	1							10	5.500
27	-	-	-	1	-	-	-	-							1	5.000
28	-	-	1	2	-	-	2	-							5	6.000
29	-	1	1	-	1	-	-	-							3	4.333
30	-	-	-	-	1	-	-	-							1	6.000
>30	-	11[c]	28[d]	20[e]	16[f]	10[g]	3[h]	4[i]	3[j]	1[k]	-	1[l]			97	5.412
total	1	101	346	336	346	337	257	202	139	104	44	23	11	2	2249	6.768
occurrences per word ($\bar{y}_x$)	3.00	22.08	11.17	9.42	6.47	5.15	4.29	3.98	3.32	2.55	1.77	4.35	1.36	1.50		

a/ x=17 b/ x=15 c/ y = 36, 47, 48, 48, 53, 58, 134, 139, 297, 305, 638.

d/ y=31, 33, 34, 35, 36, 38, 40, 41, 41, 42, 45, 46, 49, 51, 54, 63, 68, 69, 87, 102, 116, 117, 119, 131, 135, 189, 216, 292.

e/ y=31, 35, 35, 36, 37, 38, 40, 42, 44, 49, 49, 49, 58, 90, 96, 111, 112, 160, 168, 261.

f/ y=31, 31, 34, 36, 39, 41, 42, 42, 49, 52, 55, 57, 63, 76, 94, 124.

g/ y=31, 35, 36, 37, 43, 44, 60, 64, 109, 133.

h/ y= 33,36,68; i/ y=33,42,43,55; j/ y=33, 34, 44; k/ y=35; l/ y = 39.

10.3.6. Two extensive counts available to the author were not used for the investigation. One is M. Hanley's "Word Index to Ulysses" by James Joyce and the other, "The Teacher's Handbook of 30,000 words" by E. L. Thorndike and I. Lorge (1944). The former was rejected in view of the extremely unusual vocabulary of 'Ulysses'; the latter was not taken up in view of criticisms like those made in Dewey (1923, page 5 of 1950 edition) and also because even this 18 million word count cannot solve the problem raised in Section 10.5.

10.3.7. Table 10.3 presents the correlation coefficients r_{xy} for all the word-counts analysed, along with the ratios $\sum xy/n\bar{x}\bar{y}$ indicating the length of the text as a fraction of the average length expected for random matching. Both r_{xy} and $\sum xy/n\bar{x}\bar{y}$ are presented separately for the actual bivariate distributions of x and y as well as for the hypothetical distributions where matching between words and meanings is most economic and most uneconomic.

10.3.8. Since the interest lies in results based on infinitely large counts, the estimates shown in Table 10.3 seem to be inadequate, being based on relatively frequent words. Such questions will be discussed in Section 10.5; provisional estimates of r_{xy} etc. will be presented in that Section.

10.4.1. The Correlation Ratios : Table 10.4 presents the correlation ratios η_{xy} and η_{yx} for all the word-counts, separately for actual data and for most efficient and most inefficient matchings.

Table 10.3. Correlation coefficient (r_{xy}) between word-length in letters (x) and word-frequency (y) and its effect on text-length, for actual word-counts and for hypothetical most efficient/inefficient matchings between words (letter-combinations) and meanings or concepts.

word-count material		number of		correlation coefficient (r_{xy})			text-length as p.c. of of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$)		
		word-types (n)	word-tokens ($n\bar{y}$)	actual word-count	most efficient matching	most inefficient matching	actual word-count	most efficient matching	most inefficient matching
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. Miller et al (1958):	(a) content words	5180	14877	-0.201	-0.517	0.663	88.3	69.8	138.7
	(b) function words	357	21422	-0.314	-0.412	0.592	58.0	44.9	179.2
	(c) all words	5537	36299	-0.138	-0.199	0.279	62.4	46.1	175.8
2. Dewey (1923):	(a) particular words with y > 10	1026	78634	-0.226	-0.301	0.487	64.9	53.2	175.8
	(b) root words with y > 10	1132	87380	-0.233	-0.320	0.444	68.5	56.7	160.1
3. Yule (1944):	(a) nouns in Bunyan's Pilgrim's Progress, Pt. I	1005	4047	-0.167	-0.384	0.579	85.5	66.8	150.1
	(b) -do- Pt.II	1020	4016	-0.157	-0.417	0.653	87.6	67.1	151.5
	(c) -do- Mr.Badman	1030	3992	-0.147	-0.350	0.555	85.0	64.4	156.3
	(d) -do- Holy War	996	4001	-0.137	-0.342	0.530	86.5	66.3	152.2
	(e) -do- in all four works of Bunyan	2249	16063	-0.152	-0.348	0.525	82.4	59.8	160.8

Table 10.4 Correlation ratios between word-length in letters (x, and word-frequency y) for actual word-counts and for hypothetical most efficient/inefficient matching between words (letter-combinations) and meanings or concepts.

word-count material		number of word-types (n)	number of word tokens ($n\bar{y}$)	number of arrays with x fixed	number of arrays with y fixed	correlation ratios for different types of matching: actual word-counts η_{yx}	actual word-counts η_{xy}*	most efficient matching η_{yx}	most efficient matching η_{xy}*	most inefficient matching η_{yx}	most inefficient matching η_{xy}*
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Miller et al (1958):	(a) content words	5180	14877	17	53	0.246	0.233	0.961	0.835	0.988	0.923
	(b) function words	357	21422	13	110	0.427	0.545	0.905	0.948	0.993	0.989
	(c) all words	5537	36299	17	120	0.337	0.309	0.879	0.923	0.926	0.924
2. Dewey (1923)	(a) particular words with $y > 10$	1026	78634	14	172	0.424	0.391	0.902	0.975	0.959	0.992
	(b) root words with $y > 10$	1132	87380	12	183	0.439	0.388	0.905	0.951	0.974	0.980
3. Yule (1944):	(a) nouns in Bunyan's Pilgrim's Progress Pt. I	1005	4047	14	41	0.230	0.185	0.881	0.832	0.949	0.953
	(b) -do- Pt. II	1020	4016	14	47	0.198	0.224	0.856	0.821	0.974	0.948
	(c) -do- Mr. Badman	1030	3992	16	42	0.210	0.178	0.826	0.814	0.992	0.944
	(d) -do- Holy War	996	4001	13	40	0.154	0.187	0.858	0.832	0.995	0.965
	(e) -do- in all four works of Bynyan	2249	16063	15	87	0.184	0.191	0.841	0.887	0.975	0.980

* These correlation ratios were adjusted by using formula (4) in para 10.4.3.

10.4.2. One limitation of the correlation ratio is that it tends to increase with the number of arrays formed for presenting the bivariate distribution. No grouping of x and y was used in the present case. As a consequence, the number of y-arrays of x is found to be too large, sometimes well above 100. The correlation ratios η_{xy} should therefore be too high.

10.4.3. The correlation ratio η_{xy}, or rather its square, η^2_{xy}, measures the proportion of the total variability of x explained by the polynomial of (k-1)th degree passing through the points corresponding to the k array means $(y, \bar{x}_y)$. But surely a polynomial of degree 50 or 100 is extremely artificial and incorporates part of the erratic element in the scatter diagram. The 'true regression can certainly be approximated by a polynomial of much lower degree. Assuming that a 10th degree polynomial is adequate, and also that the mean square due to the 11th degree and higher order terms is equal to the residual (within array) mean square, we adjusted the observed correlation ratios η_{xy} by the following formula

$$\eta^{2\,*}_{xy} = \frac{(n-11)\,\eta^2_{xy} - (k-11)}{(n-k)} \qquad \ldots\ldots \qquad (4)$$

Here n is the number of word-types, k the number of arrays with y fixed, and η and η^* the observed and adjusted values, respectively, of the correlation ratio. Then η^* is, in a sense, an estimate of what is called the "correlation index" associated with the best-fitting 10th degree polynomial regression of x on y.

10.4.4. The values of η_{xy} presented in Table 10.4 (and also subsequently in Table 10.6) are all adjusted in the above-mentioned manner. Some η^*-values turned out negative; obviously the 11th and higher degree terms had a much smaller mean square than the residual component. The original η-values have been shown in such cases. No adjustment seemed to be necessary for η_{yx}-values, as the number of x-arrays is not at all large.

10.4.5. Observations on these correlation ratios will be made in Section 10.5 after discussing the problem of estimating the limiting values for infinitely large word-counts.

10.5.1. Limiting values and related questions : The figures in Tables 10.1, 10.3 and 10.4 suffer from certain limitations. In all cases, the ultimate interest is in the values of r_{xy} etc. in infinitely large word counts where even the rarest words of the language are represented with their true relative frequencies. As it is, the estimates given are based on the section of relatively frequent words and hence may be biased. This is particularly applicable to estimates obtained from the Dewey counts which are available in a truncated form above y=10. We now consider these questions in some detail.

10.5.2. If $p_1, p_2, \ldots, p_N$ are the probabilities of occurrence of the N words of the language, and $x_1, x_2, \ldots, x_N$ the respective lengths in terms of the number of letters, the "limiting" value of

r_{xy} will be r_{xp} where, using standard symbols,

$$r_{xp} = \frac{\frac{1}{N}\sum px - \bar{p}\,\bar{x}}{\sigma_p\,\sigma_x} = \frac{\frac{1}{N}\left(\sum px - \bar{x}\right)}{\sigma_p\,\sigma_x} \qquad \ldots\ (5)$$

If this calculation is based on words with $p > p_o$, some minimum value, the correlation coefficient may be denoted by $r(p_o)$. The ultimate interest is in r_{xp}, which may be denoted by $r(o)$.

10.5.3. The economic effect of the correlation has been measured by $\sum xy/n\,\bar{x}\,\bar{y}$. The "limiting" value of this is simply

$$e = \frac{\sum px}{\bar{x}} \qquad \ldots\ (6)$$

a simple yet interesting relation. Again, if the r.h.s. is based on words with $p > p_o$, we may denote the ratio by $e(p_o)$, so that e is denoted by $e(o)$.

10.5.4. Direct estimation of r(o), say, does not appear to be impossible at first sight. The sum $\sum px$ is merely the population mean of word-length, different occurrences of the same word being counted separately. The mean observed in a large count should be a reasonably good estimate of this. The values of N, $\bar{x}$ and σ_x can be found from the Oxford English Dictionary, say, where all words of the language may be supposed to be listed. The value of σ_p can be estimated by first estimating $\sum p^2$ by the methods given by Good [1953, Section 6].

10.5.5. The estimation of e(o) presents no additional problem.

10.5.6. But no attempt was made to estimate r(o), e(o) etc., following such direct methods. Apart from the question of time and labour, it was felt that these limiting values cannot be defined in a precise and meaningful manner.

10.5.7. There is, first of all, the disheartening observation by L.P. Ayres, repeatedly stressed by Dewey (op. cit, page 5) that beyond the first 500 or at most 1000 most frequent words of the language, no general statement about the relative frequencies of the comparatively rare words is possible; the relative frequencies depend most critically on the subjectmatter of the linguistic material. The length-frequency correlation in a general-purpose word count is the thing of real interest, but it is difficult to say what is meant by a general-purpose word count.

10.5.8. Second, it is difficult, if not impossible, to decide upon a list of all root words or particular words in any language, treating foreign words, archaic words and technical words in an objective manner.

10.5.9. Finally, so far as the class of all particular words is concerned, it seems to be futile to consider such limiting values. The Thorndike and Lorge count (1944) shows that when we examine rarer and rarer words we meet more and more frequently compounds and highly inflected words of ever increasing length. Conceptually, the number of particular words in English, exposed to the risk of being used, may be regarded as literally infinite — remember coining of new

words in various ways in 'Ulysses' by James Joyce. So the limit of $\sum px/\bar{x}$ for the class of particular words is nothing but zero; for the numerator is finite, about 4.5 letters, while the denominator is infinity.

10.5.10. The above difficulties are not very serious for the class of function words, and the estimates obtained for this class of words may perhaps be accepted as fair approximations to the limiting values.

10.5.11. Since direct methods for estimation of r(o) etc. did not seem to be promising, one indirect approach was made to the problem. Good (1953) had considered problems similar to the present one, and had pointed out that if a word occurs y times in a count of $n\bar{y}$ word occurrences, then $y/(n\bar{y})$ is not a satisfactory estimate of the probability of occurrence of that word when y itself is small [ibid, p.237]. Nevertheless, some idea of the variation of $r(p_o)$ and $e(p_o)$ with p_o may be obtained from Table 10.5 and Figures 10.1 - 10.3. These present the results of calculations carried out on all the counts[1] after excluding rare words upto different values of y.

10.5.12. We may first compare the results based on the Miller et al count (all words) with those for Dewey's count (particular words). As the number of word-tokens in the former count was 36,299 against 100,000 in Dewey's, one might compare the section of the former count with $y \geq 3$ with the section of the latter count with $y \geq 10$. Similarly sections of the Miller et al count with $y \geq 4,5,7,10$ and 18, may be treated as roughly corresponding to sections of Dewey's count of particular words with $y \geq 12, 15, 20, 30$ and 50 respectively.

1/ Yule's counts on individual works of Bunyan were excluded, and only the pooled count covering all four works was utilised for this purpose.

Table 10.5 : Effect of omitting low-frequency words on r_{xy} etc.

word-count material		frequencies included $y \geq y_o$	number of word-types (n)	number of word tokens ($n\bar{y}$)	mean of word-length in letters ($\bar{x}$)	mean of word-frequency ($\bar{y}$)	correlation coefficient (r_{xy}) actual word count	correlation coefficient (r_{xy}) most efficient matching	correlation coefficient (r_{xy}) most inefficient matching	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) actual word-count	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) most efficient matching	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) most inefficient matching
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. Miller et al (1958):	(a) content words	$y \geq 1$	5180	14877	7.326	2.872	-0.201	-0.517	+0.663	88.3	69.8	138.7
		≥ 2	2328	12025	6.871	5.165	-0.225	-0.564	0.742	89.8	74.6	133.4
		≥ 3	1421	10211	6.517	7.186	-0.222	-0.592	0.771	91.3	76.8	130.2
		≥ 4	963	8837	6.329	9.177	-0.238	-0.616	0.787	91.7	78.4	127.6
		≥ 5	720	7865	6.126	10.924	-0.230	-0.623	0.812	92.4	79.5	126.7
		≥ 6	554	7035	5.971	12.699	-0.231	-0.641	0.828	92.9	80.3	125.5
		≥ 8	351	5723	5.635	16.305	-0.212	-0.661	0.876	94.3	82.2	123.4
		≥ 11	212	4494	5.382	21.198	-0.206	-0.622	0.881	95.1	85.2	121.0
		≥ 19	83	2699	5.000	32.518	-0.208	-0.690	0.928	96.2	87.5	116.8
	(b) function words	$y \geq 1$	357	21422	5.392	60.006	-0.314	-0.412	+0.592	58.0	44.9	179.2
		≥ 2	290	21355	4.983	73.638	-0.318	-0.436	0.623	62.6	48.7	173.4
		≥ 3	260	21295	4.815	81.904	-0.319	-0.445	0.649	64.5	50.6	172.1
		≥ 4	241	21238	4.701	88.124	-0.318	-0.450	0.667	65.9	51.8	171.5
		≥ 5	228	21186	4.605	92.921	-0.313	-0.450	0.687	67.1	52.8	172.1
		≥ 6	215	21121	4.498	98.237	-0.315	-0.460	0.697	68.5	54.0	169.6
		≥ 8	198	21011	4.409	106.116	-0.320	-0.473	0.662	69.6	55.1	163.1
		≥ 11	175	20811	4.206	118.920	-0.310	-0.479	0.700	72.3	57.3	162.4
		≥ 19	141	20328	3.922	144.170	-0.289	-0.479	0.769	76.1	60.4	163.5
	(c) all words	$y \geq 1$	5537	36299	7.201	6.556	-0.138	-0.199	0.279	62.4	46.1	175.8
		≥ 2	2618	33380	6.662	12.750	-0.175	-0.242	0.360	63.3	49.3	175.4
		≥ 3	1681	31506	6.254	18.742	-0.196	-0.273	0.398	64.4	50.4	172.4
		≥ 4	1204	30075	6.003	24.979	-0.215	-0.299	0.412	64.8	51.1	167.3

(contd.)

Table 10.5. (Contd.)

word-count material		frequencies included $y \geq y_o$	number of word-types (n)	number of word tokens ($n\bar{y}$)	mean of word-length in letters ($\bar{x}$)	mean of word frequency ($\bar{y}$)	correlation coefficient (r_{xy}) actual word count	correlation coefficient (r_{xy}) most efficient matching	correlation coefficient (r_{xy}) most inefficient matching	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) actual word count	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) most efficient matching	text length as p.c. of average under random matching ($\Sigma xy/n\bar{x}\bar{y}$) most inefficient matching.
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. Miller et al (1958):	(c) all words (contd.)	$y \geq 5$	948	29051	5.761	30.645	-0.224	-0.315	+0.452	65.7	51.8	169.2
		≥ 6	769	28156	5.559	36.614	-0.234	-0.333	0.486	66.5	52.4	169.5
		≥ 8	549	26734	5.193	48.696	-0.252	-0.368	0.555	68.3	53.8	169.8
		≥ 11	387	25305	4.850	65.388	-0.263	-0.378	0.648	70.3	57.3	173.2
		≥ 19	224	23027	4.321	102.799	-0.278	-0.428	0.676	74.0	60.0	163.2
2. Dewey (1923):	(a) particular words	$y \geq 11$	1026	78634	5.552	76.641	-0.226	-0.301	0.487	64.9	53.2	175.8
		≥ 13	851	76632	5.371	90.049	-0.236	-0.297	0.512	65.7	56.7	174.5
		≥ 16	694	74451	5.157	107.278	-0.251	-0.341	0.533	66.8	55.0	170.5
		≥ 21	504	71081	4.861	141.034	-0.267	-0.370	0.603	68.5	56.3	171.1
		≥ 31	329	66758	4.453	202.912	-0.287	-0.415	0.630	71.5	58.7	162.6
		≥ 51	204	61939	4.025	303.623	-0.303	-0.470	0.693	75.3	61.7	156.4
	(b) root words	$y \geq 11$	1132	87380	5.384	77.191	-0.233	-0.320	0.444	68.5	56.7	160.1
		≥ 13	989	85740	5.300	86.693	-0.238	-0.326	0.464	68.7	57.2	160.9
		≥ 16	840	83675	5.163	99.613	-0.252	-0.345	0.478	69.4	58.1	158.1
		≥ 21	673	80674	5.010	119.872	-0.265	-0.364	0.511	69.9	58.7	158.1
		≥ 31	442	74941	4.661	169.550	-0.291	-0.408	0.545	71.6	60.2	153.3
		≥ 51	257	67688	4.113	263.377	-0.300	-0.459	0.642	75.8	63.0	151.8
3. Yule (1944):	nouns in Bunyan's works	$y \geq 1$	2249	16063	6.770	7.142	-0.152	-0.348	0.525	82.4	59.8	160.8
		≥ 2	1316	15130	6.459	11.497	-0.166	-0.391	0.590	84.8	64.4	153.8
		≥ 3	944	14386	6.249	15.239	-0.174	-0.416	0.636	86.4	67.5	149.7
		≥ 4	752	13810	6.178	18.364	-0.187	-0.425	0.664	86.7	69.7	147.4
		≥ 5	621	13286	6.090	21.395	-0.199	-0.444	0.674	87.1	71.3	143.6
		≥ 7	453	12369	5.938	27.305	-0.208	-0.463	0.706	87.8	72.9	141.3
		≥ 9	364	11705	5.810	32.157	-0.212	-0.470	0.730	88.6	74.2	140.0

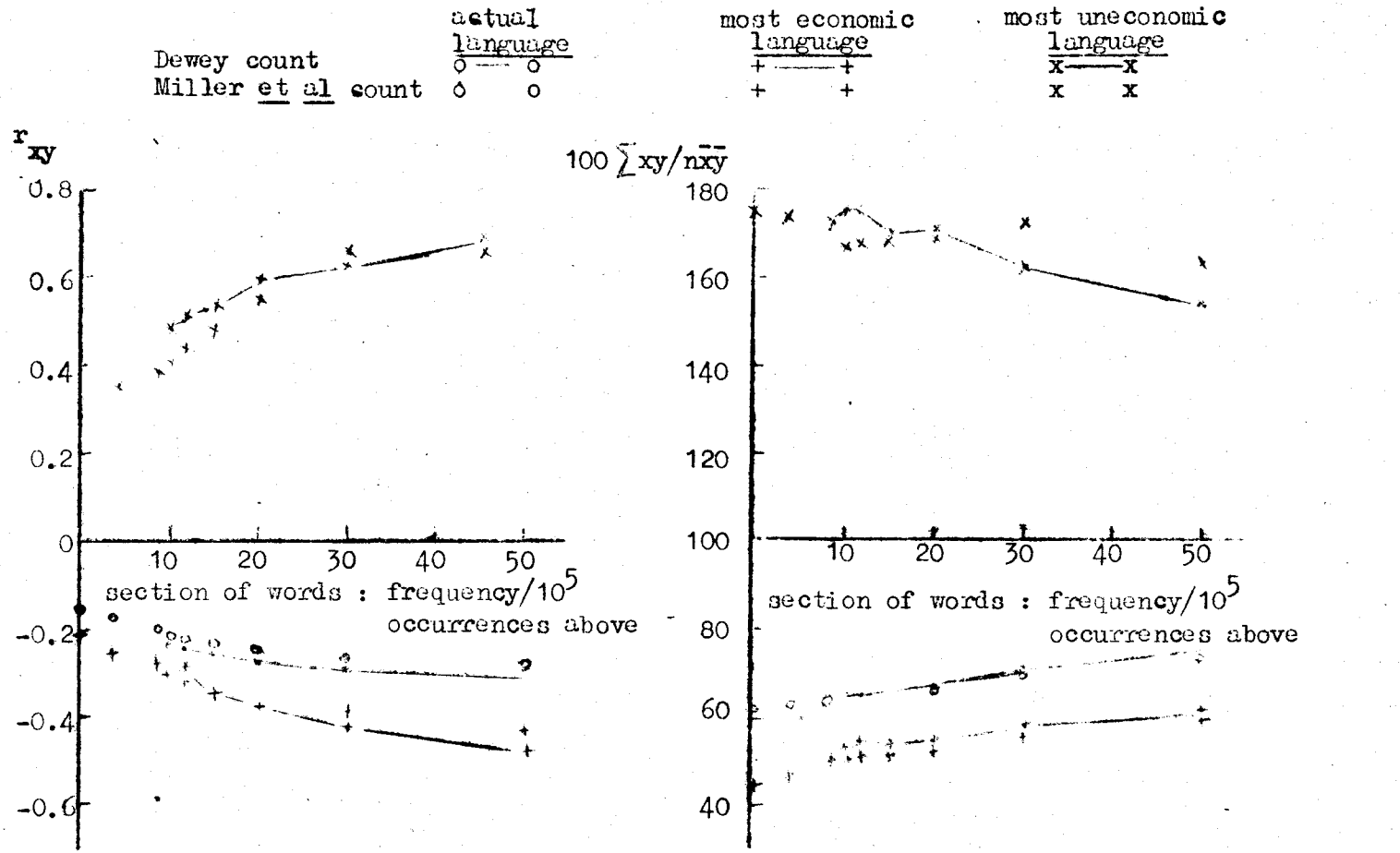


Fig.10.1 : Effect of exclusion of rare words on (a) the correlation coefficient (r_{xy}) between word-length in letters (x) and word-frequency (y) and on (b) the text-length as percentage of expected length under random matching of words and concepts ($100 \sum xy/n\bar{x}\bar{y}$) for actual language and hypothetical languages with most economic/ uneconomic matchings : based on Dewey's count of particular words and Miller et al count of 'all words'.

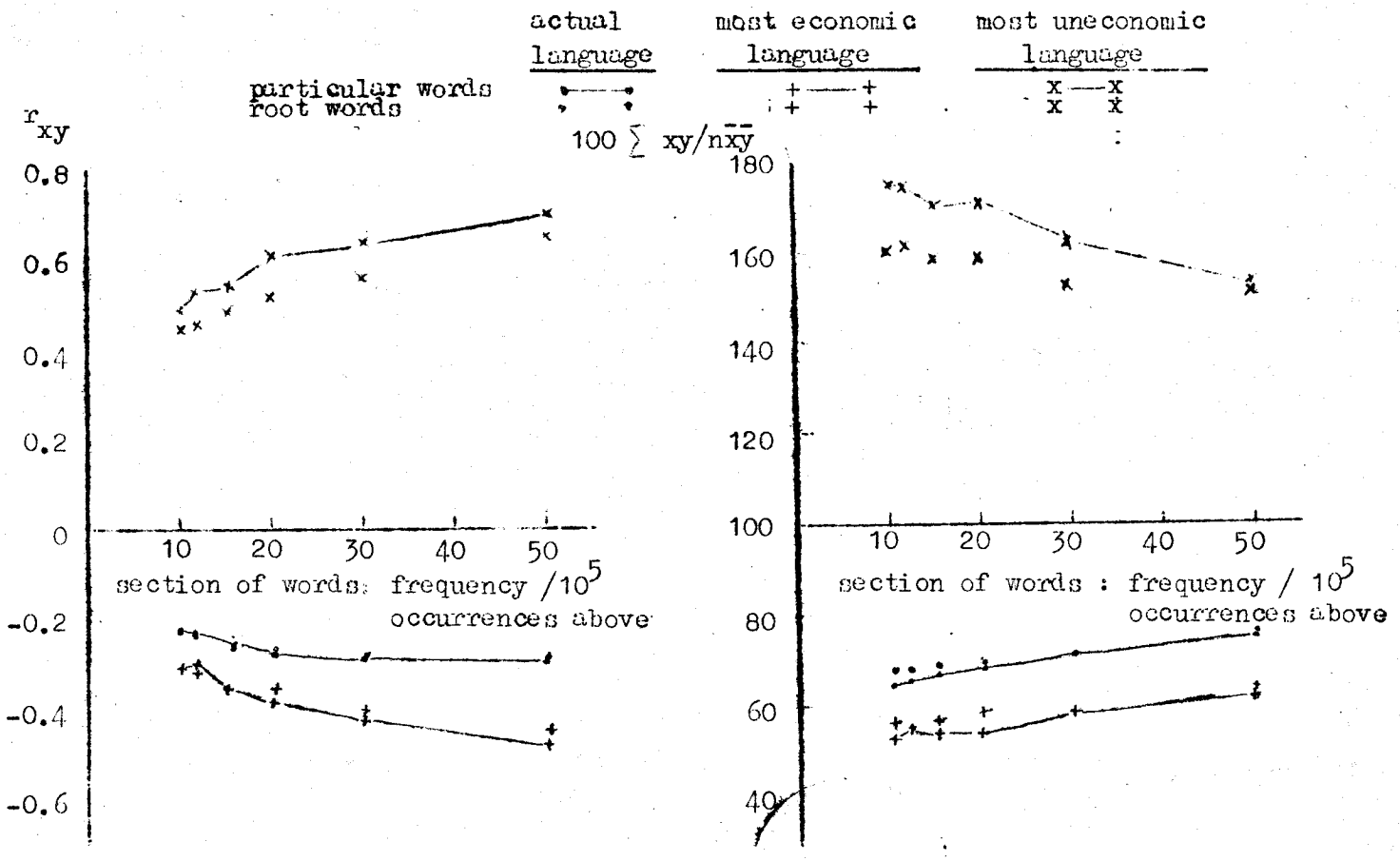


Fig. 10.2 Effect of exclusion of rare words on (a) the correlation coefficient (r_{xy}) between word-length (x) and word-frequency (y) and on (b) the text-length as percentage of expected length under random matching of words and concepts ($100 \sum xy/n\bar{x}\bar{y}$) for actual language and hypothetical languages with most economic/uneconomic matchings : based on Dewey's count of particular words and root words.

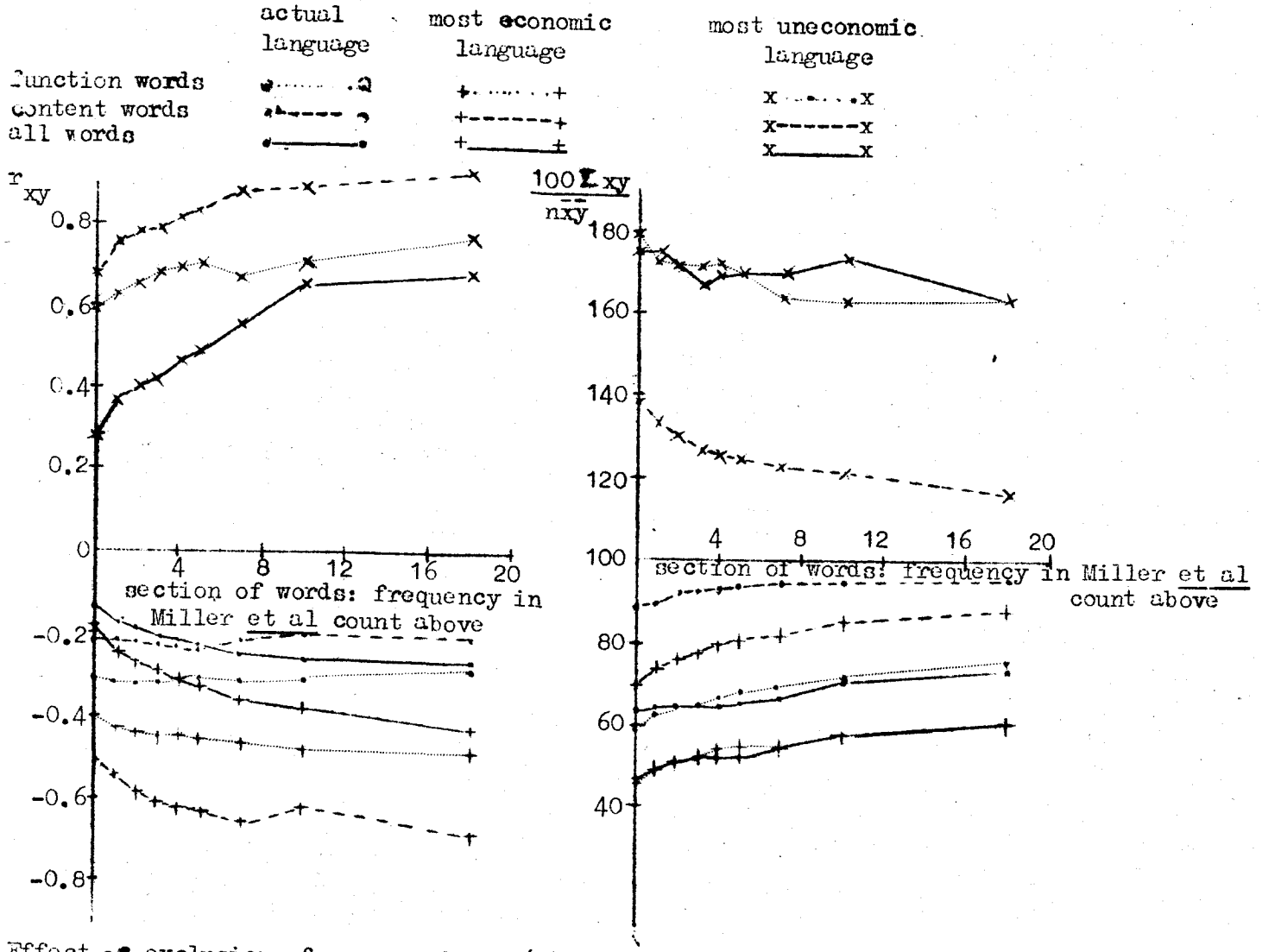


Fig.10.3: Effect of exclusion of rare words on (a) the correlation coefficient between word-length in letters (x) and word-frequency (y) and on (b) the text-length as percentage of expected length under random matching of words and concepts ($100\Sigma xy/n\bar{x}\bar{y}$) for actual language and hypothetical languages with most economic/uneconomic matchings: separately for 'function words', 'content words' and 'all words' based on Miller et al count of 36299 word-tokens.

10.5.13. Figure 10.1 shows the values of r_{xy} and $\frac{\sum xy}{nxy}$ based on corresponding sections of the two counts. The values are very much the same for the two counts. This is true for the actual values and also for the most economic and the most uneconomic matchings. This agreement is encouraging; but the estimates change systematically as rarer and rarer words are included, and the trend is such that extrapolation to get r(o) and e(o) seems to be quite unsafe. The results based on the Miller et al count upto y=1 may be adopted as the best available from this study for the class of all particular words. (The most important ones, out of these, were quoted in para 10.1.7.) But a more extensive count published complete (upto y=1) may give a somewhat different picture.

10.5.14. It is interesting to note that all three curves in Fig.10.1 show r_{xy} falling in absolute magnitude as rarer and rarer words are included. In contrast, all three curves for $\sum xy/(n\bar{x}\bar{y})$ move away from 1 as low frequency words are included.

10.5.15. As regards the root-word count of Dewey, the estimates $r(p_o)$ and $e(p_o)$ are very much similar to those for the count of particular words by Dewey [vide Fig. 10.2] and hence to those for Miller et al count of all words. The differences are usually small and unimportant. This is very fortunate. Although the agreement may not remain so good for very large counts — for infinitely large counts, we have already pointed out, it is difficult to define the measures for both classes of words — the results for the Miller et al count of particular words

upto y=1 may perhaps be taken as approximations to the unknown estimates for the root word count of Dewey upto y=3 (roughly).

10.5.16. One may now separately consider content words and function words in the count by Miller et al (1958). The graphs for r_{xy} for the actual language are very nearly horizontal for either content words or function words (vide Fig. 10.3). This means that the estimates obtained may be roughly equal to the limiting values, for either class of words. (See also para 10.5.10.) When all words are considered together, the values of r_{xy} fall in the absolute sense as rarer and rarer words are included. This is partly because as rarer words are included the proportion of content words increases and the values for content words are smaller in the absolute sense. Similar observations may be made, in a broad way, about the graphs for maximum and minimum r_{xy}.

10.5.17. As regards the estimates of $\frac{\sum xy}{nxy}$ for function words and content words, all curves in Fig. 10.3 show increasing deviations from 1 as low-frequency words are included, but the changes are not very marked in most cases.

10.5.18. The coefficient r_{xy} may be taken as about -0.20 for content words and -0.30 for function words. Miller, Newman and Friedman (1958) overemphasised this difference when they wrote that the "favoritism for some words" among the content words "is not dependent upon length". The regression of y on x is no doubt much nearer horizontal for content words [vide Figs. 4 and 5 in Miller et al (1958)] though the points for x=1 and 2 for content words should not be given

much weight. But the slope of the regression is not an indicator of the strength of correlation.

10.5.19. However, since the product CV (x) CV(y) is much larger for function words (vide Table 10.1), the economic effect of the negative r_{xy} is and can be much greater for function words than for content words [see eqn. (3)]. This is clearly shown by the curves for $\sum xy/(n\bar{x}\bar{y})$ in Fig. 10.3. The economy is smaller for content words and the two limits are also closer to 1. The all-words curve is close to that for function words, for actual, most economic and most uneconomic matchings.

10.5.20. Finally, for Yule's count on nouns in four works of John Bunyan, the values of r_{xy} show a steady decrease, in the absolute sense, as rarer words are included. So the limiting value for an indefinitely large count cannot be estimated from the observed count. The best available estimate is -0.152, the value for the whole count, upto y=1. This is quite close to the best available estimate for all root words or all particular words. But the minimum and maximum attainable values are here larger in the absolute sense; the minimum is about -0.35 instead of -0.2, and the maximum about 0.5 instead of 0.28. The maximum and minimum values also decrease in the absolute sense as rarer words are included.

10.5.21. The values of $\sum xy/n\bar{x}\bar{y}$ also show systematic movement as low-frequency words are included: The figures move away from 1 for actual language as well as for the most economic or uneconomic matching.

The best available estimates are 82.4%, 59.8% and 160.8% respectively. These seem to be nearer 1 than those for counts of particular or root words.

10.5.22. Table 10.6 shows the effect of exclusion of rare words on the correlation ratios. The estimates for the complete material are also repeated for the sake of convenience.

10.5.23. Both correlation ratios are substantially the same for comparable sections of the three counts of all words — Miller et al, Dewey (particular words) and Dewey (root words). The correlation ratios fall (but the fall is marked only for η_{xy}) when words with y between 11 and 50 (per 100,000) are excluded. The maximum seems to be reached with $y \geq 11$. But below this, only the Miller et al count is available, and the estimates therefrom decrease rapidly as rarer and rarer words are included. The best (i.e., most reasonable) estimates are those for the complete Miller et al count, viz., η_{yx}=0.337 and η_{xy} = 0.309. But even these may be considerably different from limiting values.

10.5.24 For content words, η_{yx} does not show much change as rarer words are included, and the limiting value may not be far from 0.25, the value found for the entire Miller et al count. For function words, on the other hand, the η_{yx} - values increase appreciably when low-frequency are included, but for reasons stated in para 10.5.10, the estimate 0.43 for the whole Miller et al count may not far from the limiting value. The corresponding estimates for η_{xy} might be taken as 0.23 for content words and 0.55 for function words. Non-linearity of regression, indicated by $\eta^2 - r^2$ is appreciable for function words and all words but not for content words.

Table 10.6: Effect of omitting low-frequency words on the correlation ratios

word-count material	frequencies included: $y \geq y_o$	number of word types (n)	number of word tokens ($n\bar{y}$)	no. of arrays: x fixed	no. of arrays: y fixed	correlation ratios[1]: η_{yx}	correlation ratios[1]: η_{xy}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Miller et al (1958):							
(a) content words	$y \geq 1$	5180	14877	17	53	0.246	0.233
	≥ 2	2328	12025	17	52	0.274	0.253
	≥ 4	963	8837	13	50	0.284	0.246
	≥ 8	351	5723	12	46	0.286	0.125
	≥ 19	83	2699	9	35	0.254	0.570[2]
(b) function words	$y \geq 1$	357	21422	13	110	0.427	0.545
	≥ 2	290	21355	12	109	0.413	0.508
	≥ 4	241	21238	11	107	0.401	0.475
	≥ 8	198	21011	10	103	0.382	0.501
	≥ 19	141	20328	10	92	0.340	0.396
(c) all words	$y \geq 1$	5537	36299	17	120	0.337	0.309
	≥ 2	2618	33380	17	119	0.385	0.354
	≥ 4	1204	30075	13	117	0.394	0.379
	≥ 8	549	26734	12	113	0.381	0.351
	≥ 19	224	23027	10	102	0.356	0.233
2. Dewey (1923) :							
(a) particular words	$y \geq 11$	1026	78634	14	172	0.424	0.391
	≥ 21	504	71081	12	162	0.418	0.393
	≥ 51	204	61939	9	132	0.377	0.134
(b) root words	$y \geq 11$	1132	87380	12	183	0.439	0.388
	≥ 21	673	80674	12	173	0.433	0.424
	≥ 51	257	67688	12	143	0.391	0.394
3. Yule (1944) : nouns in four works of Bunyan	$y \geq 1$	2249	16063	15	87	0.184	0.191
	≥ 2	1316	15130	14	86	0.208	0.175
	≥ 3	944	14386	13	85	0.234	0.149
	≥ 4	752	13810	12	84	0.247	0.165
	≥ 5	621	13286	11	83	0.274	0.192
	≥ 7	453	12369	11	81	0.267	0.197
	≥ 9	364	11705	11	79	0.272	0.094

1/ η_{xy} - values have been adjusted using formula (4) in para 10.4.3.

2/ Adjusted η^2_{xy} became negative in this case; so original value has been shown.

10.5.25. For Yule's count on nouns, the best estimate of η_{xy} is 0.19 and this may be close to truth, but the best estimate for η_{yx}, viz., 0.184 may be rather wide off the mark. Nonlinearity of regression is not at all conspicuous for this class of words also.

10.6.1. Some observations on the joint distribution of x and y : Out of the two regressions, that of y on x seems to be the more tractable one, if word counts of varying sizes are considered, but even this is not independent of the word-count size.

10.6.2. Let the number of distinct words (word-types) be $n = \sum_{x,y} f_{xy}$ and the number of word-occurrences (word-tokens) $n' = \sum_{x,y} y f_{xy}$. Then the number of occurrences of all x-lettered words, i.e., $\sum_{y} y f_{xy}$ may be regarded as an unbiased estimate of $n'\pi'_x$, where π'_x is the probability of some x-lettered word occurring at any particular position. Let π_x be the proportion of x-lettered words 'expected' in a list of n distinct words found in a count of n' words. Then

$$\bar{y}_x = \frac{\sum_y y f_{xy}}{\sum_y f_{xy}} \simeq \frac{n'\pi'_x}{n\pi_x} = \frac{n'}{n} \cdot \frac{\pi'_x}{\pi_x} \qquad \ldots\ldots (7)$$

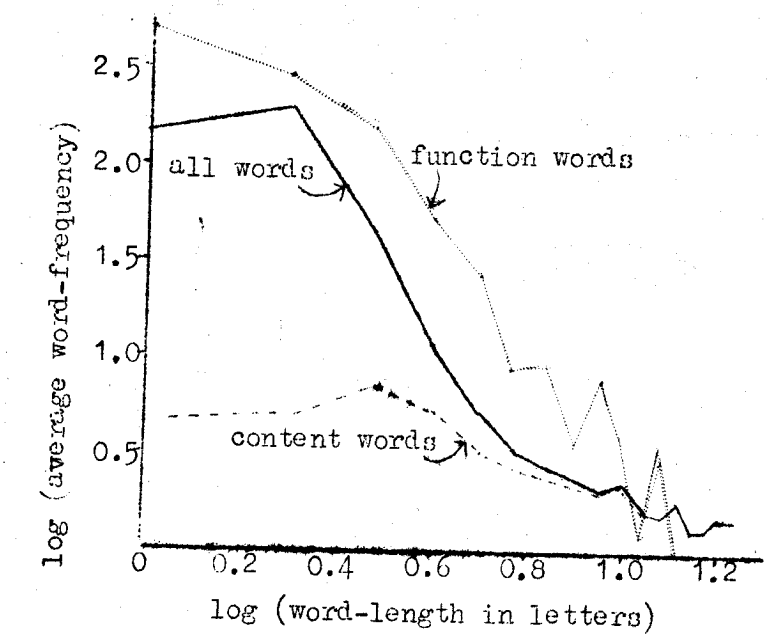
10.6.3. Now n does not increase proportionately to n' (vide Chap.2, Section 2.2). But what is really important, the proportions π_x vary systematically as n' rises : the distribution gradually shifts towards higher x-values. It follows that the $\bar{y}_x$ - values found for one value of n' would not even tend to be proportional to those for a different value of n' .

10.6.4. In view of this, only a rough examination was made of the form of this regression separately for function words, content words and all words, utilising the Miller et al count[1]. Fig. 10.4 presents the results. The upper graph shows that the regression is approximately linear on double-log scale when content words and function words are considered separately — one may ignore the points for x=1 which are based on 2 words for function words and on 5 words for content words — but when all words are studied without any subdivision into classes, the regression is appreciably curved, being convex to the x-axis. This contradicts the finding of Herdan (1958a), who showed that the regression of y on x (x = no. of letters or phonemes) was approximately $\bar{y}_x = ax^{-2.4}$[2]. But Herdan's material comprised the 738 most frequent words accounting for 76,054 occurrences in the count of nearly 80,000 conversational word-tokens by French, Carter and Koenig (1930). Obviously, the regression based on the most frequent words would give a distorted estimate of the true regression.

10.6.5. The lower part of Fig. 10.4 shows that the regression of $\log\log \bar{y}_x$ on $\log x$ is approximately linear for all words; and the curvature is not marked for content words also (the point for x = 2 is based on 12 words). For function words, however, the curvature seems to be appreciable.

1/ Section 10.6 is only a preliminary report on small-scale studies.

2/ This was demonstrated indirectly by first showing that length distributions of both vocabulary and occurrences were lognormal, and then calling to aid the moment distribution property of the log-normal distribution.



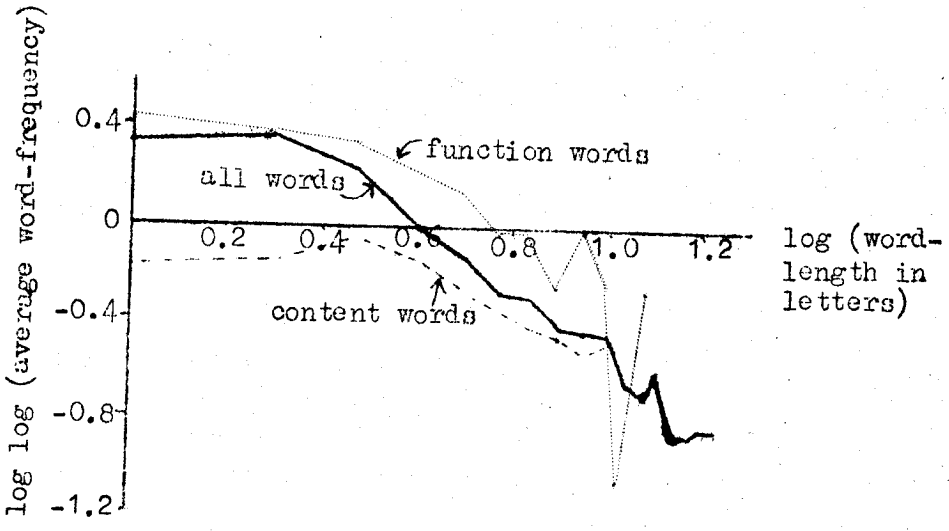


Fig. 10.4: Regression of word-frequency (y) on word-length in letters (x), based on the count by Miller et al (1958) on English prose : above - log y against log x, below-log log y against log x.

10.6.6. The conditional distribution of y given x : Considering words of a specified length x, the conditional distribution of words according to frequency y may perhaps obey the rank-frequency relation observed for the marginal distribution of y[1]. The Miller et al count of all words showed that the conformity is close for low values of x (e.g. 4), but as x increases to 9 or 10, the rank-frequency curve becomes flatter and flatter. Also there is some tendency to top-downward concavity for certain values of x (e.g. 7).

10.6.7. In order to see the position for a really large count, we examined the conditional distributions of all four-lettered and all seven-lettered words in the 'Ulysses' using the word-index by Hanley already referred to. Fig. 10.5 shows the rank-frequency relation for the 4729 seven-letter words, and also for the 2166 four-letter words. (The conditional distributions are shown in Table 10.7.) The former is sensibly linear; but the latter is not, even after excluding the first 50 words, and the "top-downward concavity" cannot be explained by the "informality" of the writing etc., as done by Zipf (1949, pp. 121-122). Also, the slope is a little steeper for the four-letter words.[2]

1/ The rank-frequency relation implies Pareto distribution of word-frequencies (Simon, 1953). For a discussion of this relation, see Chapter 2, Section 2.3 and also Appendix 5 which presents some evidence from Bengali prose.

2/ The rank-frequency relation may not be linear on double-log scale for all sizes of counts, and especially for small ones and the slope may increase with size of count (vide Zipf, 1949, p. 121).

Table 10.7 Distribution of all four-letter and seven-letter words in "Ulysses" according to the number of occurrences, based on Hanley's "Word Index to Ulysses".

number of occurrences	number of four-letter words	number of seven-letter words	number of occurrences	number of four-letter words	number of seven-letter words	number of occurrences	number of four-letter words	number of seven-letter words	number of occurrences	number of four-letter words	number of seven-letter words
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	714	2437	18	12	9	35	7	1	52	3	
2	327	817	19	16	13	36	7	5	53	3	2
3	164	417	20	12	10	37	2	5	54	2	
4	123	207	21	10	10	38	1	1	55	2	1
5	104	186	22	8	5	39	4	2	56	3	3
6	67	107	23	9	6	40	3	1	57	1	1
7	57	87	24	7	7	41	6	1	58	2	
8	45	65	25	6	8	42	6	5	59	-	
9	36	61	26	8	5	43	2	2	60	3	1
10	32	34	27	10	5	44	6	2	61	1	
11	34	31	28	7	8	45	1	2	62	4	4
12	22	25	29	4	3	46	1	3	63	2	
13	20	24	30	3	1	47	1	1	64	1	
14	23	18	31	4	2	48	4	3	65	4	
15	15	19	32	6	3	49	3		66	2	
16	13	14	33	10	2	50	1		67	1	
17	14	11	34	2	3	51	2		68 & more	131[a]	23[b]
									total	2166	4729

a/ 68,68,68,69,71,72,74,74,75,75,76,76,77,78,78,78,79,80,80,81,81,82,82,83,84,85,85,86,87,88,88,88,90,94,95,100,102,102,103.103,103,104,104,105,105,106,107,108,110,113,114,117,117,120,120,121,123,124,125,125,126,127,128,129,132,133,133,140,140,144,147,150,151,152,157,165,165,166,~~166~~170,171,172,174,178,178,183,184,185,193,195,196,197,205,211,218,218,224,224,239,248,252,270,272,277,306,312,318,318,324,330,337,343,360,376,440,448,470,480,483,510,556,569,674,688,729,897,1010,1081,1209,2506,3082.

b/ 72,72,74,81,82,82,90,97,102,113,116,117,118,119,121,127,133,138,169,213,216,245,500.

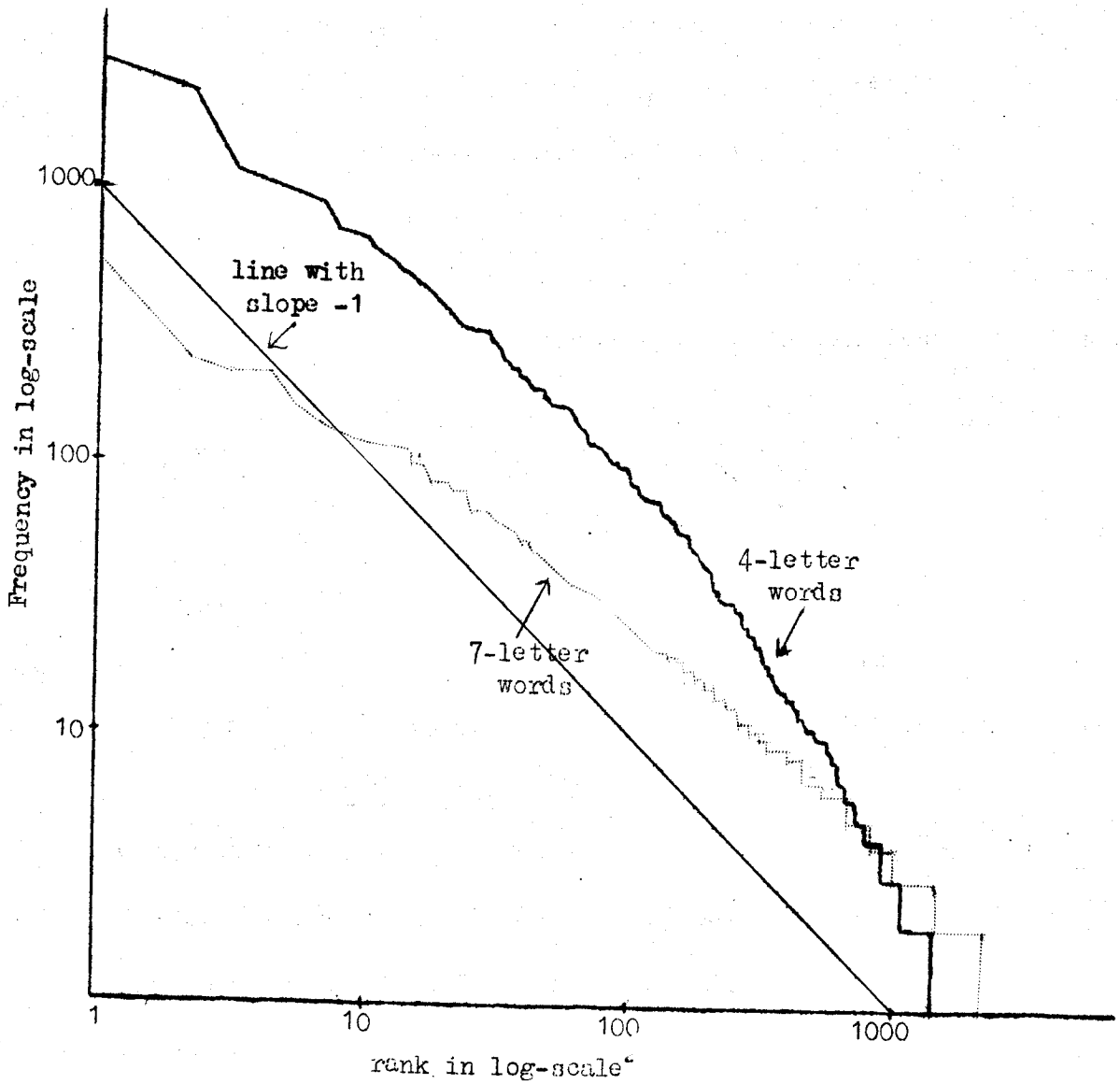


Fig.10.5: Showing rank-frequency relation for 4729 seven-letter words and 2166 four-letter words in 'Ulysses', based on Hanley's 'Word Index to Ulysses'.

10.7.1. Concluding Observations : Miller and Newman (1958) /vide also Miller, Newman and Friedman, 1958/ tried to simplify a stochastic process model of Mandelbrot (1954) which sought to explain the well-known rank-frequency relation observed in word-counts. They pointed out that two "essentially non-linguistic effects", viz, (i) the relative unlikelihood of longer strings of letters (words), and (ii) the greater variety of long words, suffice to produce a negative correlation between word-length and word-frequency. And these two non-linguistic effects are ensured if the occurrence of letters, especially spaces, is reasonably haphazard, if not exactly random.

10.7.2. This statistical explanation of the negative correlation is extremely interesting, but it is not clear how far this negative correlation can explain the rank-frequency relation. For example, why should the frequencies fall in the harmonic progression ? And why should there be so much variation in frequency among words of the same specified length ? (Vide supra)

10.7.3. The F-R relation does not agree with the r-f relation satisfactorily as claimed by Miller and Newman (1958) and Miller, Newman and Friedman, (1958). It is and has to be flatter, as the rank correlation between x and y is far from perfect. The logarithmic graph (Fig.2) of Miller and Newman (1958) understates the differences /See also Fig.3 of Miller, Newman and Friedman (1958)/.[1]

1/ These Figures show the F-R relation by plotting logarithm of average frequency against logarithm of average rank. Strictly, averages of logarithms should have been plotted along both axes.

10.7.4. However, Mandelbrot (1955) himself admitted that his model was too simple for natural languages. Miller, Newman and Friedman (1958) also admitted that their 'statistical' explanation was inadequate as it failed to predict the results when content and function words are studied separately.

10.7.5. The present study is confined to word counts on written English. It would be interesting to carry out similar analysis for other languages. It would also be interesting to analyse conversational Engligh, where the correlation might be more pronounced. The material thrown up by French, Carter and Koenig (1930) would be useful for this purpose.

10.7.6. Again, word-length has been measured here by the number of letters comprising the word. From the linguistic point of view, the number of phonemes or even syllables would probably be superior, at least for analysing conversational material. "English Pronouncing Dictionaries" like that by Daniel Jones (1963) or that by Hornby, Gatenby and Wakefield (1948) might be used for such purposes.

10.7.7. Apart from these, the present study suffers from the limitation that, no estimates could be given of the limiting values of r_{xy} etc., for infinitely large word counts [vide Section 10.5 for detailed discussion].

Chapter 11: Intervals between successive occurrences of the same word

11.1.1. Introduction : Suppose one numbers the different word-positions of a long text serially in the natural reading order and notes the positions $i_1, i_2, \ldots, i_f$ occupied by a given word occurring f times in the text. Then $x_1 = i_2 - i_1$, $x_2 = i_3 - i_2$, ..., $x_{f-1} = i_f - i_{f-1}$ are the f-1 intervals between successive occurrences of the word[1]. The distribution (of lengths) of such intervals will be studied in this chapter with material from English and Bengali texts. Some work will also be reported on the randomness of the series of intervals $x_1, x_2, \ldots, x_{f-1}$.

11.1.2. The work of Zipf (1945 , 1949, pp.40-54) is mostly based on the intervals for rare words in James Joyce's "Ulysses". Zipf utilised Hanley's (1937) Word-Index for the purpose. Since each word gave only a few intervals, Zipf pooled the intervals for all words occurring the same number of times (f) for preparing the frequency distributions of intervals. Let n_x denote the number of intervals of length x in any such distribution (x=1,2,3,...). By adopting dubious statistical procedures (see infra) Zipf found the approximate empirical relation

$$n_x = a_f \, x^{-b_f} \qquad (1)$$

for fourteen values of f between 5 and 24.[2] Here a_f and b_f are constants; actually b_f was close to 1 in most cases.

1/ Perhaps i_1 might be regarded as another and the first interval.

2/ The Hanley Word-Index gives page and line reference to each occurrence of the word only for words occurring upto 24 times in "Ulysses".

11.1.3. Zipf (1949, ibid) also examined the separate distributions of $x_1, x_2, \ldots, x_{f-1}$ and concluded that they were more or less the same for each f.

11.1.4. Herdan (1956, pp. 122-5) reported an elegant study on the intervals for the very frequent Russian grammar form К (='to'etc). The frequency distribution of 243 intervals could be very well fitted by the exponential distribution. Since the exponential is but the continuous analogue of the discrete geometric distribution, Herdan's result can be put in the form

$$n_x = ab^x \qquad (2)$$

where a, b are constants.

11.1.5. Neither Zipf nor Herdan used any goodness of fit or other statistical tests, but in Herdan's case the fit was evidently very good.

11.1.6. The above results provided the point of departure for the present study. First of all, the findings of Zipf and Herdan seemed to be conflicting: While eqn. (1) is linear on double-logarithmic scale, eqn. (2) becomes linear on the semi-logarithmic scale. But Zipf's result is based on rare English words occurring only 5 to 24 times in "Ulysses" containing 260,430 words, and Herdan's on a single very frequent grammar word in Russian. Possibly, eqns. (1) and (2) are the two limiting cases of a more general form covering all classes of words. Second, Zipf's procedure of establishing eqn. (1) was clearly defective and his material seemed to merit a more rigorous examination. Third,

it was necessary to investigate whether eqn.(2) holds for other frequent words or grammar forms. Fourth, randomness of the series of intervals seemed to deserve closer scrutiny. It is possible that $x_1, x_2, \ldots, x_{f-1}$ are autocorrelated even when they are identically distributed.

11.1.7. The following section is devoted to the re-examination of some of the interval distributions studied by Zipf (1949, ibid). It is found that eqn. (1) is not at all suitable for the material as claimed by Zipf. The geometric distribution [eqn.(2)] is better, but even this is far from adequate. Section 11.3 studies the interval-distributions for each of four very frequent grammar words (viz., 'the', 'to', 'and' and 'of') in Chapters 1-9 of the English novel, "Pride and Prejudice" by Jane Austen[1]. The geometric distribution seemed to the clear choice; but actual fits were not fully satisfactory. The Kolmogorov distances were about 6 to 8 per cent, and were, on the whole, significant. However, the geometric law was more successful for these frequent grammar words than for the rare words in "Ulysses". Section 11.4 examines the interval distributions for the words 'se' (he/she), and 'sei' (that) and the 'se' class of words in the Bengali novel, "Gora", by Tagore. Here again, eqn.(1) failed completely, but the geometric distribution seemed to give a first approximation. As in Section 11.3, the K-distances were significant, but not large in the absolute sense. In general, for all the distributions in Sections 11.2 — 11.4, shorter intervals were relatively more frequent than could be expected from the

1/ The counts needed in Sections 11.3 and 11.4 were carried out by the present author.

geometric law. Finally, Section 11.5 applies nonparametric tests and other methods and demonstrates that the series of intervals are very nearly random.

11.1.8. Interval-counts are extremely exacting and time-consuming. The work reported in this chapter is therefore on a very modest scale. But as stated in paragraphs 1.3.31 - 1.3.35 of Chapter 1, such studies are of great theoretical interest, as they throw much light on the fundamental question of applicability of probability models to texts or samples of speech considered as sequences of words. The geometric distribution of intervals [eqn.(2)], it may be pointed out, suggests the simplest probability model conceivable. Suppose each word-position is filled up at random by either the given word or by some other word. If the probability of the given word occurring is a constant p for all word-positions and if the different positions are filled in independently, the probability distribution of intervals will be of the geometric form

$$p_x = pq^{x-1} \qquad (x = 1, 2, 3, \ldots..) \qquad (3)$$

11.2.1. Re-examination of Zipf's material : We shall first describe the methods adopted by Zipf and point out his mistakes. He measured intervals in pages, subtracting the page number of each occurrence from the page number of the next occurrence. To avoid the interval zero, all the intervals were increased by 1. [Zipf admitted in a footnote (ibid, pp. 40-41) that the addition of $\frac{1}{2}$ would probably

have been better.[7] He then plotted the logarithms of the number of intervals n_x of length x against the logarithm of x, and felt that the relationship was linear. He then fitted eqn.(1) after double-logarithmic transformation by ordinary least squares[1/]. Actually, he took eqn.(1) in the form $n_x^p \; x$ = constant.

11.2.2. The greatest defect of Zipf's procedure was that he used the part of the distribution upto x = 21 (upto 50 in one case), thereby ignoring a fair proportion of observations in the upper tail. Zipf was conscious of this defect (ibid, p.44), but wrongly asserted that the effects were unimportant. Since the interest was centred upon the linearity or otherwise of the relationship between log n_x and log x, it was not advisable to study only a part of the range of log x. Fig. 11.1 below shows that the relationship is far from linear. Apparently, Zipf was troubled by the fact that for many of the larger x-values, n_x was found to be zero, which created difficulties in his procedure of examination and fitting. But this difficulty could easily be obviated by standard statistical methods based on the cumulative frequency distribution.

11.2.3. The discrete probability distribution defined by the probability

$$p_x = a/x^b \qquad (x = 1,2,3,.....) \qquad (4)$$

(where a,b are positive constants and p_x the probability of the integer x) may be called the Zeta distribution, in view of its connection with

1/ The use of ordinary unweighted least squares was not quite appropriate, but similar use is often made, for instance, in fitting the Pareto law to grouped income distributions (Bhattacharya and Mukherjee, 1965).

the Riemannian Zeta function (Whittaker and Watson, 1958). Zipf's relation [eqn.(1)] is clearly equivalent to this zeta distribution. Obviously, $b > 1$; otherwise $\sum_{1}^{\infty} p_x$ does not converge. Also, $a = 1/\zeta(b)$ in order that $\sum_{1}^{\infty} p_x = 1$. This, therefore, is a uniparametric distribution.

11.2.4. The mean of the distribution exists if $b > 2$ and in that case is given by $\sum_{1}^{\infty} x p_x = \zeta(b-1)/\zeta(b)$. For variance to exist, b must be greater than 3. Zipf did not notice such points. He concluded that p is of the order of 1, which means b is of the order of 1; but b must be clearly greater than 1. In the present case, one may even lay down that $b > 2$. A zeta distribution with b near 1 is a priori unsuitable for fitting the observed distribution of intervals.[1/]

11.2.5. Zipf was, however, justified to a great extent in the use of the page as the unit of interval length. This is seen as follows. Consider eqn. (4). If the x-values are arranged into groups, such that the rth group comprises the integers $(r-1)K + 1$ to rK $(r=1,2,\ldots)$, the total probability of the rth group is, using the continuous Pareto

---

1/ Yule (1944, p. 55) made a similar criticism of Zipf's law of distribution of words according to frequencies of occurrence (x) in a sufficiently long text, viz., $p_x \sim a/x^b$, with $b \sim 2$. Apart from disagreeing that the fit was good even for Zipf's own data, Yule commented that this distribution does not have a finite mean unless $b > 2$, not to speak of variance. In the opinion of the present author, an infinite mean for this distribution may be quite meaningful, for in an infinitely large word-count, the number of word-types may have a finite limiting value, so that the average frequency per word may tend to ∞. [Vide Feller, 1957, Section X.1 , on probability distributions with infinite expectation.]

approximation (see next para)

$$\text{Prob.}\left\{(r-1)\,K+1 \leq x \leq rK\right\} \sim \frac{a}{b-1}\,K^{-b+1}\left[(r-1)^{-b+1}-r^{-b+1}\right]$$

$$\sim aK^{-b+1}\,r^{-b}$$

ignoring terms of higher order in $1/r$. So the form of the distribution is hardly affected by such grouping except for small values of r. If the distribution of intervals in terms of words follow eqn.(4), the distribution in terms of pages would also have the same form except for small values of r. As Zipf was considering rare words, the frequencies at small values of r were not very important.

11.2.6. We come now to the methods proposed to be adopted in our re-examination. For reasons already stated, the graphical examination should preferably be based on cumulative frequencies. Eqn. (4) does not lead to a tractable distribution function $F(x) = \sum_{1}^{x} p_r$. But the zeta distribution can be approximated by the continuous Pareto law with frequency function $f(x) = ax^{-b}$, defined from $x = \frac{1}{2}$ (approx.) to ∞ : the probability $\int_{x-\frac{1}{2}}^{x+\frac{1}{2}} ax^{-b}dx$ between $x-\frac{1}{2}$ and $x+\frac{1}{2}$ according to this Pareto law would be very nearly equal to the probability $p_x = ax^{-b}$ of the discrete zeta distribution for x=1,2,3,... The approximation will improve with increasing x. For large x we have

$$1-F(x) = \int_{x+\frac{1}{2}}^{\infty} ax^{-b}dx = \frac{a}{b-1}\left(x+\tfrac{1}{2}\right)^{1-b} \sim \alpha\, x^{-\beta} \qquad (5)$$

where $\alpha, \beta > 0$. Thus, $\log\left[1-F(x)\right]$ and hence log (cumulated frequency above x) should bear a linear relationship to log x unless x

is small. This gives a graphical method of examining the suitability of the zeta distribution[1]. The cumulation of frequencies obviates the difficulty created by the zero values of n_x in the upper tail of the distribution.

11.2.7. Given a random sample of observations $x_1, x_2, \ldots, x_n$ from the zeta distribution [eqn.(4)] the maximum likelihood estimate of b is the solution of

$$f'(b) / f(b) = -\log G \qquad (6)$$

where $f'(b) = d f(b)/db$, and G is the geometric mean of the x's. We did not, however, actually fit the zeta distribution to any data, since the graphical test indicated very poor fit in all cases.

11.2.8. Zipf examined the interval distributions for words occurring f times in "Ulysses", separately for f= 5,6,10,12,15,16,17,18,19,20,21, 22,23 and 24. The present re-examination covers three representative values of f, viz., 5, 15 and 24. In each case, the Hanley Index was scanned completely and words occurring f times picked out. Table 2-3 of Zipf (op.cit, p.43) shows that 906, 96 and 34 words occur 5, 15 and 24 times, respectively, in "Ulysses"; but in the present investigation, we could find 898, 97 and 33 words respectively[2]. The small excesses or deficits cannot have mattered appreciably in the conclusions.

1/ This is essentially the popular method of examining whether the Pareto distribution can fit a given grouped distribution of (say) income.

2/ We may mention, for example, that the mimeographed copy of the Index available to the author actually showed 34 words occurring 24 times each, but for one word the page references to occurrences were not printed.

11.2.9. The intervals were determined in the same manner as described by Zipf, adding 1 to the results of subtraction, to avoid zero intervals. The three frequency distributions of intervals are shown in Table 11.1, in a somewhat grouped form in the upper sections of the range.

11.2.10. We begin with the relatively frequent words, occurring f =24 times. There are 759 intervals of 33 words. Fig. 11.1 is drawn in the manner of Zipf and shows log n_x against log x, but here the whole range of values of x, upto 190, has been considered. The problem of zero frequencies has been tackled by grouping 5 or 10 consecutive values[1] of x and plotting the average n_x against average x on double-log scale[2]. It is clear how wrong Zipf was in confining his attention to x =1 through 21; the relationship is clearly curved (concave to the horizontal axis) above this range. The zeta distribution is thus not suitable for f=24.

11.2.11. It may be shown that if the graph of log n_x against log x is approximately linear, so would be the graph of log $\lceil 1-F(x) \rceil$ against log x, and vice versa. Also, concavity of either graph implies concavity of the other. For f=5 and f=15, we plotted log $\lceil 1-F(x) \rceil$ against log x. In both cases, the graph was appreciably concave to the horizontal axis. It was evident beyond doubt that the Zipf relation [eqn.(1)] i.e., the zeta distribution [eqn. (4)] was not at all satisfactory for Zipf's material.

1/ Five for x between 51 and 100, and ten for x between 101 and 190.

2/ The arithmetic averages were used, instead of the geometric averages, but the difference obviously could not matter.

Table 11.1: Distribution of intervals according to length in terms of pages between successive occurrences of the same word, separately for 898 words each occurring 5 times, 97 words each occurring 15 times and 33 words each occurring 24 times in "Ulysses".

length (no. of pages	no. of intervals of words each occurring		
	5 times	15 times	24 times
(1)	(2)	(3)	(4)
1	295	172	69
2	121	87	55
3	74	57	43
4	48	39	28
5	54	30	24
6	37	27	24
7	46	32	21
8	35	29	18
9	31	20	13
10	22	23	8
11	33	26	23
12	40	17	14
13	35	17	11
14	27	15	20
15	22	16	15
16	38	13	13
17	29	15	10
18	24	21	9
19	21	18	6
20	29	9	7
21	30	10	4
22	31	11	8
23	20	15	10
24	23	10	14
25	26	9	12
26	11	12	6
27	25	16	7
28	21	13	7
29	17	10	11
30	15	11	5
31	14	6	3
32	12	12	10
33	16	9	5
34	14	10	8
35	12	4	4
36	25	7	6
37	19	8	5
38	15	8	3
39	19	7	4
40	13	11	5
41	9	6	6
42	10	2	7
43	24	10	5
44	26	9	4
45	10	6	3
46	15	8	4
47	12	7	6
48	16	7	6
49	10	9	4
50	17	8	5
51-55	76	27	12
56-60	64	30	9
61-65	64	20	10
66-70	58	20	10
71-75	44	29	9
76-80	68	26	4

(contd.)

Table 11.1: (contd.)

length (no.of pages)	no. of intervals of words each occurring 5 times	15 times	24 times	length (no.of pages)	no. of intervals of words each occurring 5 times	15 times	24 times
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
81-85	58	21	6	241-250	36	4	1
86-90	51	17	6	251-260	27	2	
91-95	58	23	8	261-270	34	3	
96-100	50	8	8	271-280	35	3	
101-105	56	13	6	281-290	24	3	
106-110	40	14	7	291-300	30	2	
111-115	35	14	5	301-310	23	1	
116-120	35	14	6	311-320	21	1	
121-125	41	5	3	321-330	21	4	
126-130	33	12	4	331-340	20	-	
131--135	35	9	3	341-350	25	-	
136-140	38	4	2	351-360	22	-	
141-145	23	9	3	361-370	26	-	
146-150	31	5	2	371-380	25	-	
151-155	32	6	1	381-390	14	-	
156-160	39	5	1	391-400	7	-	
161-165	35	6	1	401-425	31	2	
166-170	33	4	5	426-450	21	-	
171-175	34	4	-	451-475	20	-	
176-180	29	4	2	476-500	9	-	
181-185	41	2	3	501-525	8	-	
186-190	33	5	1	526-550	12	-	
191-195	43	3	-	551-575	4	-	
196-200	24	6	-	576-600	5	-	
201-210	48	3	1	601-625	2	-	
211-220	52	3	2	626-650	1	-	
221-230	38	6	-	651-675	2	-	
231-240	55	2	-	total	3592	1358	759

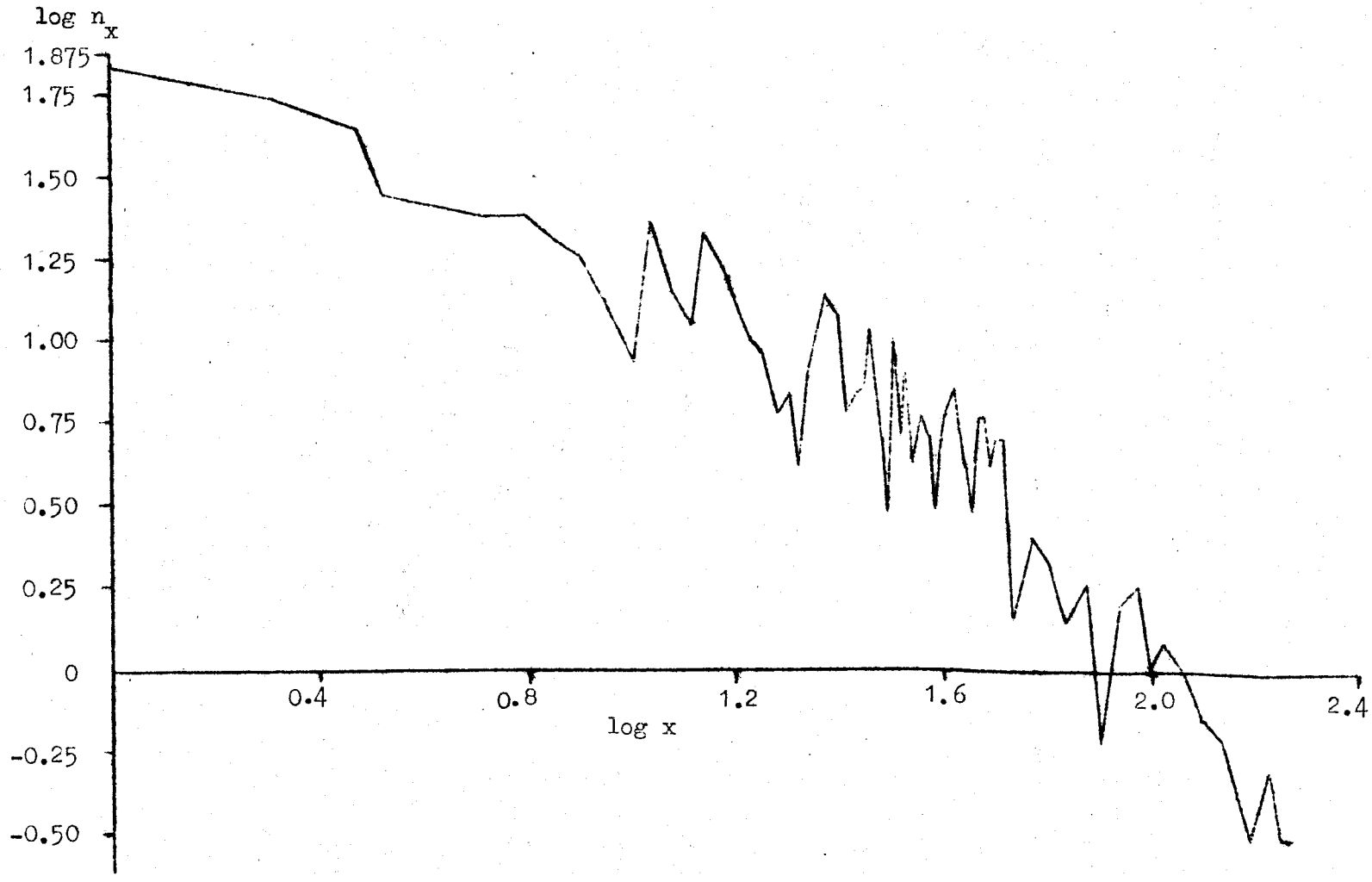


Fig.11.1: Showing frequencies (n_x) of intervals of different lengths x in terms of pages among the 759 intervals for 33 words occurring 24 times each in "Ulysses" [Vide Table 11.1.]

11.2.12. Attempts were made to fit the geometric distribution to these interval distributions of Table 11.1. The graphs showing $-\log\{1-F(x)\}$ against x seemed to be of the desired type[1]. But detailed calculations showed serious divergence between observed and expected frequencies. It was then realised that the small deviations of the graph from straight lines through the origin were more serious than believed at first sight.

11.2.13. If $x_1, x_2, \ldots, x_n$ are independent observations from the geometric population $p_x = pq^{x-1}$ $(x = 1,2,\ldots)$, the maximum likelihood estimate of p is given by $\hat{p} = 1/\bar{x}$, where $\bar{x}$ is the sample arithmetic mean. This result was used to estimate the parameter p (and hence $q = 1-p$). Calculation of expected frequencies was made on that basis.

11.2.14. For f=15, $\bar{x}$ was 44.09, so that $\hat{p} = 0.02268$. The following shows the observed F(x) and expected $\hat{F}(x)$ for selected values of x:

x :	1	2	5	10	20	51	101	301
F(x):	0.127	0.191	0.284	0.380	0.503	0.705	0.868	0.994
$\hat{F}(x)$:	0.022	0.045	0.108	0.205	0.368	0.690	0.901	0.999

The observed proportion of intervals was too high for x=1 and 2, or more generally, upto x=10 or 11, roughly.[2] But the observed proportions

---

1/ Since $F(x) = \sum_{1}^{x} pq^{r-1} = 1-q^x$, the graph of $-\log\{1-F(x)\}$ against x should be a straight line through origin.

2/ Evidently, the divergence is highly significant by the Kolmogorov test. The test is presumably somewhat conservative here, since a parameter has been estimated from the sample. The deviations from simple random sampling are not really serious; also the discreteness of the variate can be safely ignored.

declined more rapidly and above x=10, the agreement was more or less satisfactory. This suggested that the geometric distribution might be fitted to the interval distribution truncated at x=11, where observed $F(x) = 39.9\%$.

11.2.15. When the geometric distribution $p_x = pq^{x-1}$ $(x=1,2,.....)$ is truncated to have $x \geq k$, the probabilities become $p'_x = pq^{x-k}$ and the distribution function $F'(x) = 1 - q^{x+1-k}$. Therefore the graph of $-\log [1 - F'(x)]$ against $(x - k + 1)$ resembles a straight line rising from the origin. In the present case, the graphical test was very well satisfied after the truncation at $x = 11$ (vide Fig. 11.2). (The corresponding graph before truncation showed a slight but clear decline in the slope over the lowest ranges of x.) It seemed unnecessary to compare the observed and expected frequencies for the truncated distribution.

11.2.16. We refrain from quoting details here, but the position seemed to be roughly the same for f=24; the geometric distribution failed mainly because the observed frequencies were rather too high for small values of x, upto 10, roughly speaking.

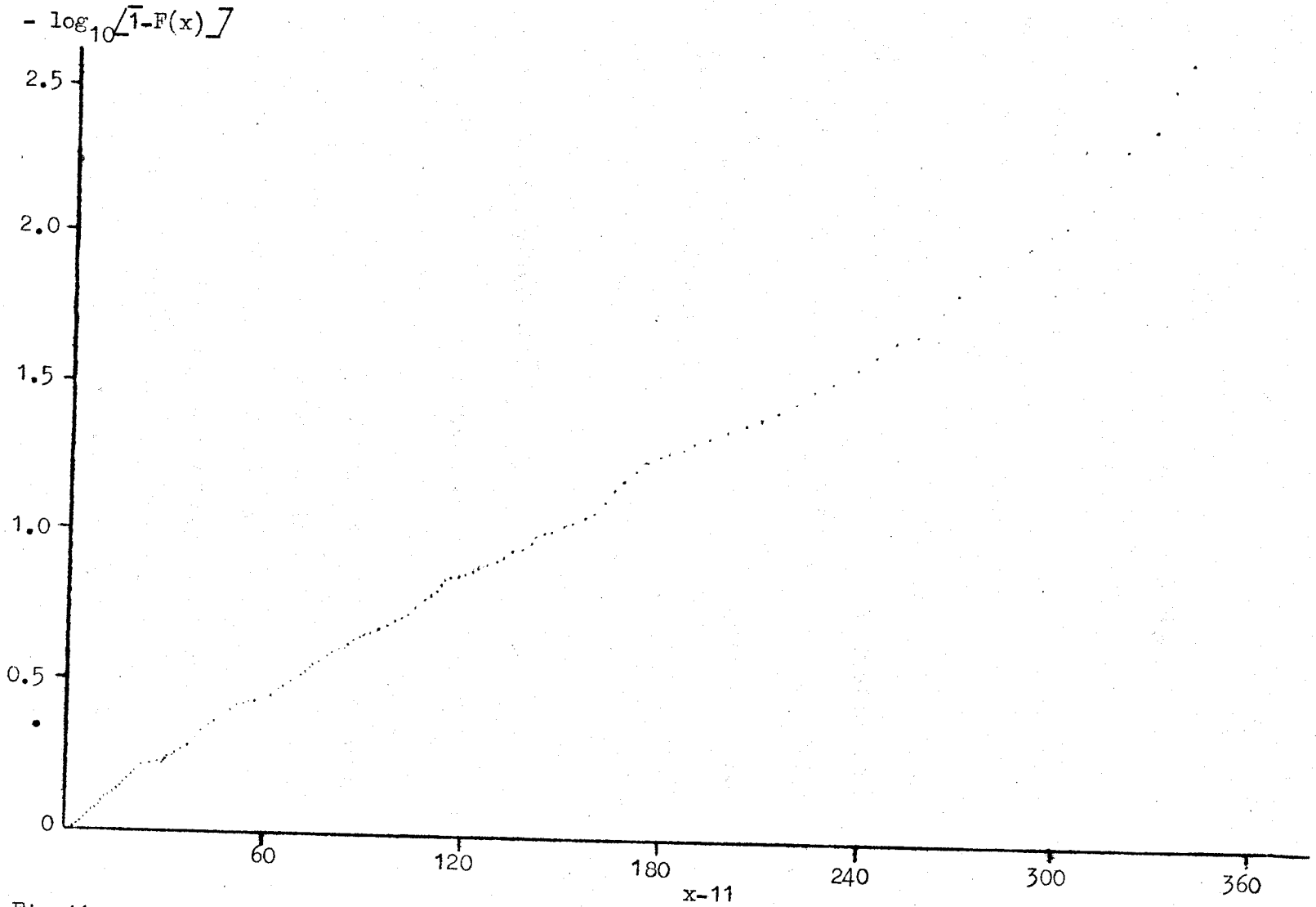


Fig.11.2 : Showing - $\log_{10} [\bar{1}-F(x)]$ against x-11 for the distribution of intervals in terms of pages (x) truncated below x=12 : 97 words occurring 15 times each.

11.2.17. For f=5, however, the intervals were much longer, on the average, /and the region of poor fit extended upto about x=35, approximately. Since $\bar{x}$ = 109.18, $\hat{p}$ came out as 0.009159. Observed and expected values of the cumulative proportion were as follows for selected values of x :

x :	1	6	11	21	51	101	201	501
F(x):	0.082	0.179	0.221	0.303	0.445	0.610	0.806	0.989
$\hat{F}$ (x):	0.009	0.054	0.096	0.176	0.375	0.605	0.846	0.990

Detailed examination showed that there were too many intervals in the observed data with length upto about 35 pages, and too few in the range 36-100, roughly. In the range 101-300, also, the observed frequencies were a little on the lower side. The Kolmogorov statistic was as large as 12.94% (at x=16), which is very highly significant.

11.2.18. To sum up, then, the geometric distribution failed for the distributions of Table 11.1, primarily because the frequencies of short intervals were much too high compared with the geometric model[1/]. This might be due to the nature of these rare words — which is quite different from the frequent grammar words studied in the two following sections, where the geometric law would prove to be more satisfactory. The occurrence of these rare words might be more heavily influenced by the context, and if the context gradually changes, it is quite natural to find too many short intervals than predicted from the geometric model.

1/ This was also seen from the graphs showing log n_x against x which were clearly convex to the horizontal axis, especially over the lower ranges of x.

11.2.19. The opposite signs of curvature in the two graphs, (i) showing log n_x against log x, and (ii) showing log n_x against x, suggested that the lognormal distribution might give a reasonable fit to the interval distributions. Of course, the integral length x would be taken as corresponding to the interval (x-1 to x) of the lognormal distribution. Fig. 11.3 shows that the expectation is not fulfilled, that is, the lognormal fit is not quite satisfactory.

11.3.1. Four grammar words of English : Four very frequent grammar words, viz., 'the', 'to', 'and' and 'of', were chosen for the investigation. According to Dewey's Count (Dewey, 1923), these are the four most frequent words in modern English prose[1/]. For each of these words, separately, a count was made of all intervals between successive occurrences in the first nine chapters of the English novel, "Pride and Prejudice", by Jane Austen. Intervals were measured in terms of the number of intervening words. (This is less by 1 than the measure discussed in the two preceding sections.) The interval preceding the first occurrence was also included [cf. para 11.1.1 and footnote thereto].

11.3.2. Table 11.2 shows the length distribution of intervals, separately for the four words.

1/ In Dewey's count of 100,000 word-tokens, 'the' occurred 7310 times, 'of' 3998 times, 'and' 3280 times, and 'to' 2924 times. But the relative frequencies in individual works may be quite different, as in "Pride and Prejudice" [Vide also Appendix II of Hanley].

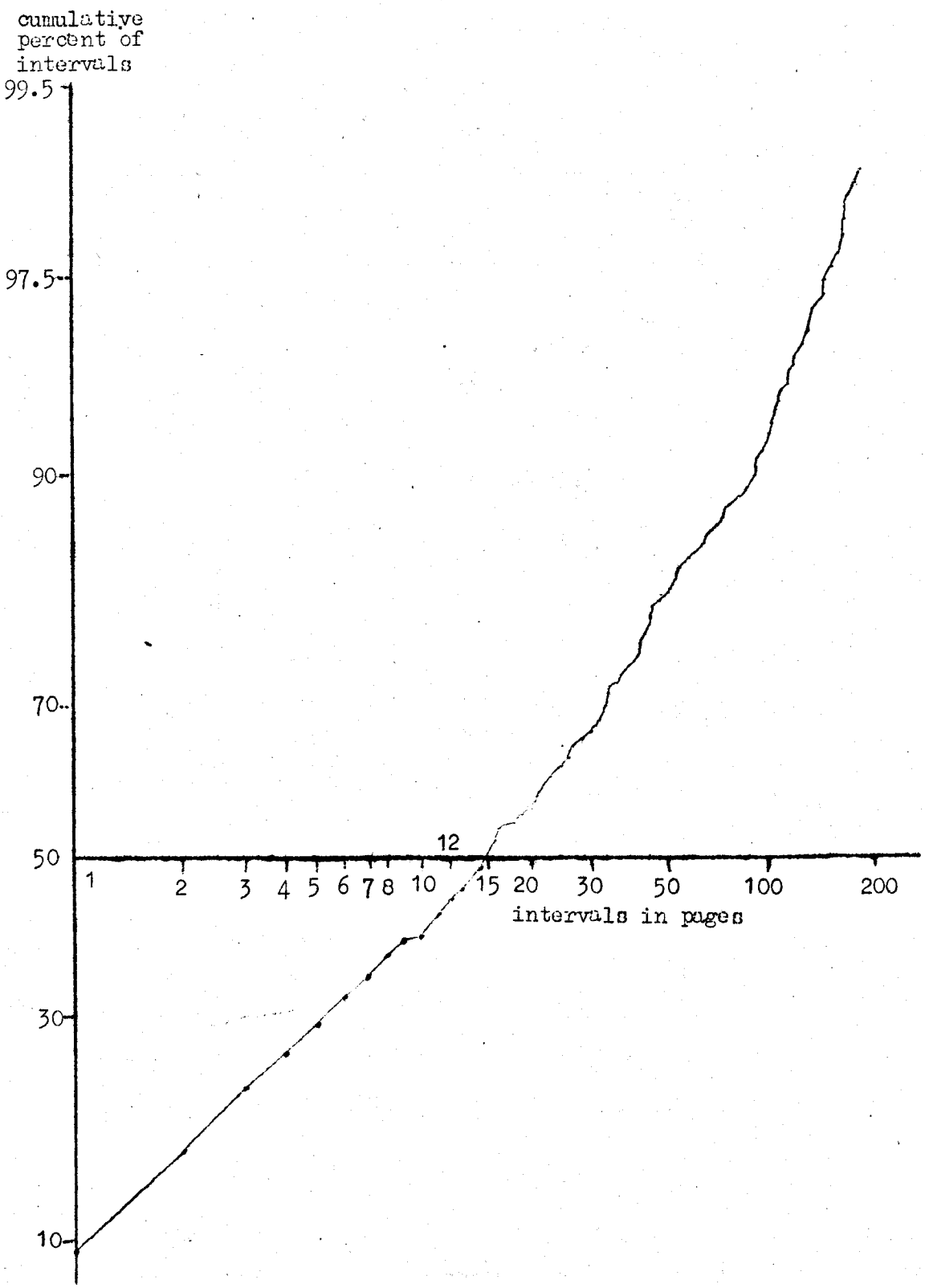


Fig. 11.3: Showing ogive on log-probit scale for the distribution of 759 intervals in terms of pages for 33 words occurring 24 times each in "Ulysses" (vide Table 11.1).

Table 11.2: Frequency distribution of intervals according to length in terms of intervening words between successive occurrences of each of four words ("the", "to", "and" and "of") in "Pride and Prejudice" (Chapters 1-9).

length (no.of words)	number of intervals				length (no.of words)	number of intervals			
	"the"	"to"	"and"	"of"		"the"	"to"	"and"	"of"
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1	7	10	6	1	28	4	3	9	4
2	29	10	5	5	29	6	9	2	6
3	20	9	7	13	30	4	3	3	2
4	18	12	3	5	31	5	7	2	5
5	13	9	11	13	32	5	9	3	5
6	16	10	9	12	33	1	7	5	4
7	17	9	8	7	34	7	3	7	5
8	9	12	8	12	35	1	6	4	6
9	14	12	4	12	36	3	2	4	1
10	13	8	5	9	37	4	5	3	3
11	14	17	4	10	38	6	7	2	2
12	10	10	16	7	39	5	3	5	7
13	11	9	5	12	40	2	6	8	3
14	19	14	7	4	41	2	5	3	4
15	8	8	11	6	42	2	2	6	9
16	3	9	11	5	43	3	2	4	1
17	6	11	5	3	44	4	3	4	2
18	7	4	6	6	45	2	1	3	5
19	5	10	14	5	46	3	2	5	1
20	11	10	3	5	47	4	4	4	2
21	6	8	9	7	48	4	1	1	3
22	4	3	2	6	49	3	3	1	2
23	5	4	7	6	50	4	4	3	4
24	4	5	2	3	51	3	1	1	1
25	10	5	6	7	52	2	3	1	
26	7	5	8	4	53	3	5	7	2
27	5	7	4	2	54	1	3	6	1

(contd.)

Table 11.2: (contd.)

length (no.of words)	number of intervals "the"	"to"	"and"	"of"	length (no.of words)	number of intervals "the"	"to"	"and"	"of"
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
55	1	1	4		79	1	2	1	2
56		3		1	80		1	2	
57	2	3	2	2	81		1	2	
58	2		3	2	82	2			
59	2		2	1	83	1	1	1	2
60	2	3	1	5	84	1	1		1
61	1	1	2		85	1		1	2
62		4	2	1	86		2		
63	1	1	1	1	87	1			
64	1		3	1	88		2	1	1
65	1	1	1	4	89			1	
66		2	2	2	90			1	
67		4	1		91				2
68		2	2	2	92				1
69	2	1	1	2	93		1		
70	1		2		94	2		1	1
71	1	1		2	95			2	1
72	1			1	96	1	1		
73			2	1	97	1		1	
74	3	1			98				
75	1	3	2	2	99		3		1
76	2	1	2	1	100				2
77		2	2	1	>100	23(a)	16(b)	17(c)	23(d)
78	2	1	2	1	total	444	415	362	349

(a) 101,101,106,112,112,114,119,121,121,129,132,135,138,141,147,152,152, 172,178,182,205,213 & 232.
(b) 104,109,110,112,114,119,120,120,124,139,142,147,148,157,158 & 183.
(c) 106,106,107,111,114,121,122,132,141,141,162,162,164,166,171,180 &186.
(d) 102,102,104,104,105,105,105,113,118,122,126,131,151,151,154,161,168, 175,208,225,250,257 & 381.

11.3.3. As already stated, interval-counts of this type are very exacting and time-consuming. A few occurrences are likely to be missed and this would introduce a definite bias in the estimated frequency distribution. A check is available when the same text is used for counts on more than one word. Thus, all the four distributions of Table 11.2 lead to very nearly the same estimate, 13,353, of the total number of words in Chapters 1-9 of "Pride and Prejudice". But this check does not safeguard against omissions.

11.3.4. Figs. 11.4(a) - (b) show $-\log [1-F(x)]$ against $x+1$ for all the four distributions. As before, $F(x)$ is the cumulative proportion of intervals with length upto x. These graphs looked sensibly close to lines passing through the origin. This suggested that the geometric distribution would successfully fit the observed data.

11.3.5. Since 'the' occurred 444 times in the text of 13353 words, the estimate $\hat{p}$ was taken as 444/13353= 0.03325, and $\hat{q}$ as 0.96675.[1/] Similar estimates were obtained for the other three words. The 'expected' distribution was obtained by using the relation $1 - \hat{F}(x) = \hat{q}^{x+1}$. The observed and 'expected' distributions are shown in terms of cumulative percentages in Table 11.3. Only selected values of x are covered, however. The fitted distributions are also shown by straight lines in Figs. 11.4(a) - (b).

1/ This estimate is slightly different from that mentioned in para 11.2.13, since 13353 includes the number of words in Chapter 9 of the novel after the last occurrence of 'the' in the said Chapter, and the number of words upto and including the first occurrence of 'the' in Chapter 1.

Table 11.3 : Cumulative frequency distribution of intervals according to length obtained from the frequency distributions of Table 11.2, along with the 'expected' distributions based on the geometric law.

length (no. of words)*	observed and "expected" cumulative percentages of intervals							
	'the'		'to'		'and'		'of'	
	obs.	exp.	obs.	exp.	obs.	exp.	obs.	exp.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	0	3.33	0	3.11	0	2.71	0	2.61
1	1.58	6.54	2.41	6.12	1.66	5.35	0.29	5.16
2	8.11	9.65	4.82	9.04	3.04	7.91	1.72	7.64
3	12.61	12.65	6.99	11.86	4.97	10.41	5.44	10.05
4	16.67	15.56	9.88	14.60	5.80	12.84	6.88	12.40
5	19.59	18.36	12.05	17.26	8.84	15.80	10.60	14.69
6	23.20	21.08	14.46	19.83	11.33	17.50	14.04	16.91
7	27.03	23.70	16.63	22.32	13.54	19.74	16.05	19.09
8	29.05	26.24	19.52	24.73	15.75	21.91	19.48	21.21
9	32.02	28.69	22.41	27.07	16.85	24.03	22.92	23.27
10	35.14	31.06	24.34	29.34	18.23	26.09	25.50	25.27
12	40.54	35.57	30.84	33.66	23.76	30.04	30.37	29.13
14	47.30	39.78	36.39	37.72	27.07	33.78	34.96	32.78
16	49.77	43.72	40.48	41.53	33.15	37.33	38.11	36.25
18	52.70	47.40	44.10	45.11	36.19	40.68	40.69	39.54
20	56.31	50.84	48.92	48.47	40.88	43.85	43.55	42.66
22	58.56	54.06	51.57	51.62	43.92	46.86	47.28	45.62
24	60.59	57.06	53.73	54.58	46.41	49.70	49.86	48.42
26	64.41	59.87	56.14	57.36	50.28	52.39	53.01	51.08
28	66.44	62.49	58.55	59.97	53.87	54.94	54.73	53.61
30	68.69	64.95	61.45	62.42	55.25	57.34	57.02	56.00
35	72.97	70.40	69.16	67.91	61.05	62.82	64.18	61.46
40	77.48	75.00	74.70	72.60	67.13	67.60	68.77	66.24

(contd.)

Table 11.3: (contd.)

length (no. of words)*	observed and "expected" cumulative percentages of intervals							
	'the'		'to'		'and'		'of'	
	obs.	exp.	obs.	exp.	obs.	exp.	obs.	exp.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
45	80.40	78.89	77.83	76.60	72.65	71.76	74.78	70.43
50	84.46	82.18	81.20	80.02	76.52	75.38	78.22	74.09
55	86.71	84.95	84.34	82.93	81.77	78.54	79.37	77.31
60	88.51	87.29	86.51	85.43	83.98	81.30	82.52	80.12
65	89.41	89.27	88.19	87.55	86.46	83.70	84.53	82.59
70	90.09	90.94	90.36	89.37	88.67	85.79	86.25	84.75
75	91.44	92.35	91.57	90.92	89.78	87.62	87.97	86.64
80	92.57	93.54	93.25	92.25	92.27	89.21	89.40	88.30
85	93.69	94.54	93.98	93.38	93.37	90.59	90.83	89.75
90	93.92	95.39	94.94	94.35	94.20	91.80	91.12	91.02
95	94.37	96.11	95.18	95.17	95.03	92.85	92.55	92.13
100	94.82	96.71	96.14	95.88	95.30	93.77	93.41	93.11
110	95.50	97.66	96.87	96.99	96.13	95.27	95.42	94.71
120	96.40	98.33	98.07	97.81	96.69	96.40	95.99	95.94
130	97.07	98.81	98.31	98.40	97.24	97.27	96.56	96.89
150	98.20	99.39	99.28	99.15	98.07	98.42	96.85	98.17
175	98.87	99.74	99.76	99.61	99.45	99.21	98.57	99.05
200	99.32	99.89	100.00	99.82	100.00	99.60	98.57	99.51
225	99.77	99.95	100.00	99.92	100.00	99.80	99.14	99.75
250	100.00	99.98	100.00	99.96	100.00	99.90	99.43	99.87
no. of intervals	444		415		362		349	
K-distance (%)	7.52		5.69		7.86		5.92	

* only selected values shown in the table.

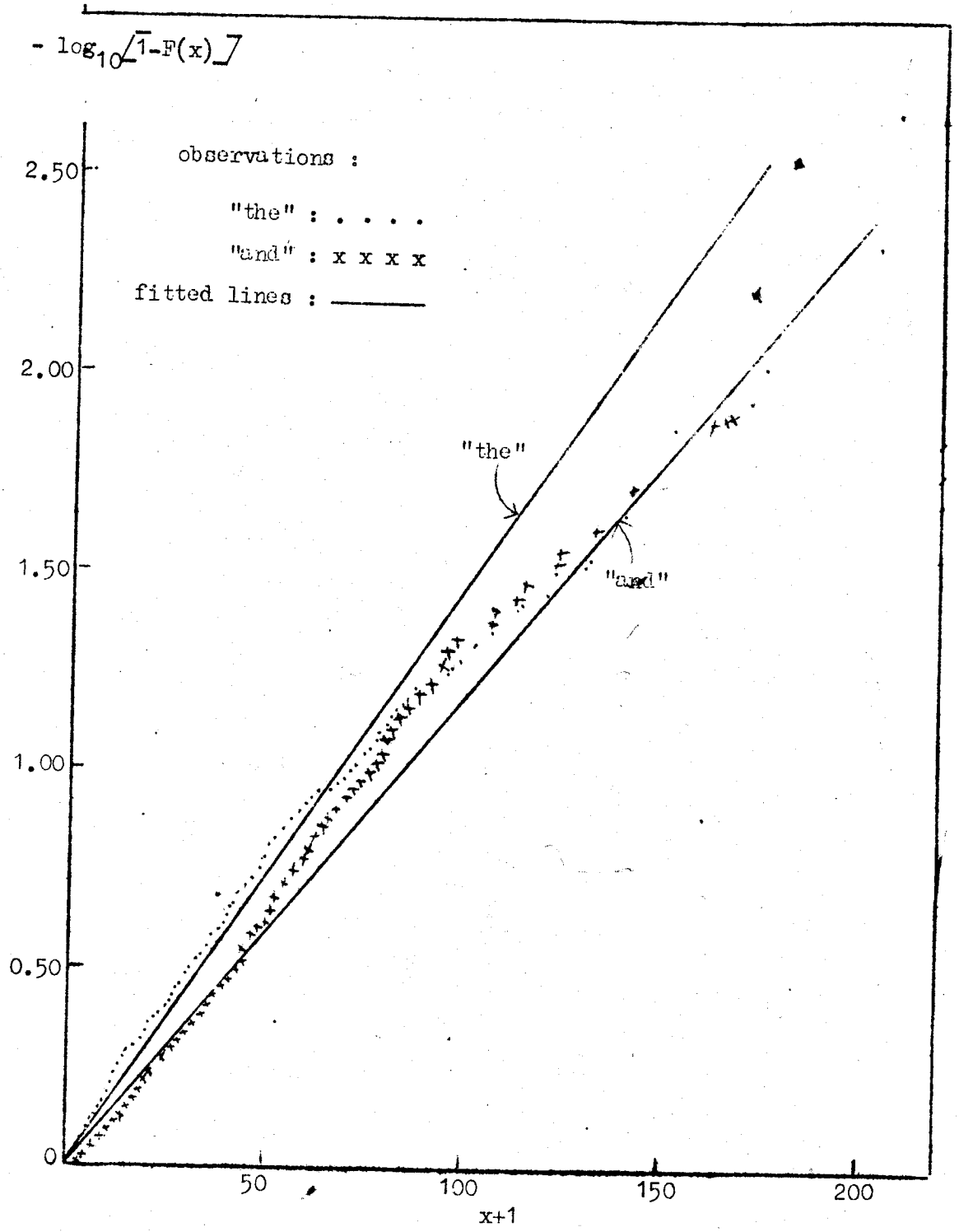


Fig.11.4(a): Showing $-\log_{10}[1-F(x)]$ against x+1, where x is the number of words intervening between successive occurrences of the word "the" or "and" and F(x) the cumulative proportion of intervals with length upto x: based on Chapters 1-9 of "Pride and Prejudice" by Jane Austen.

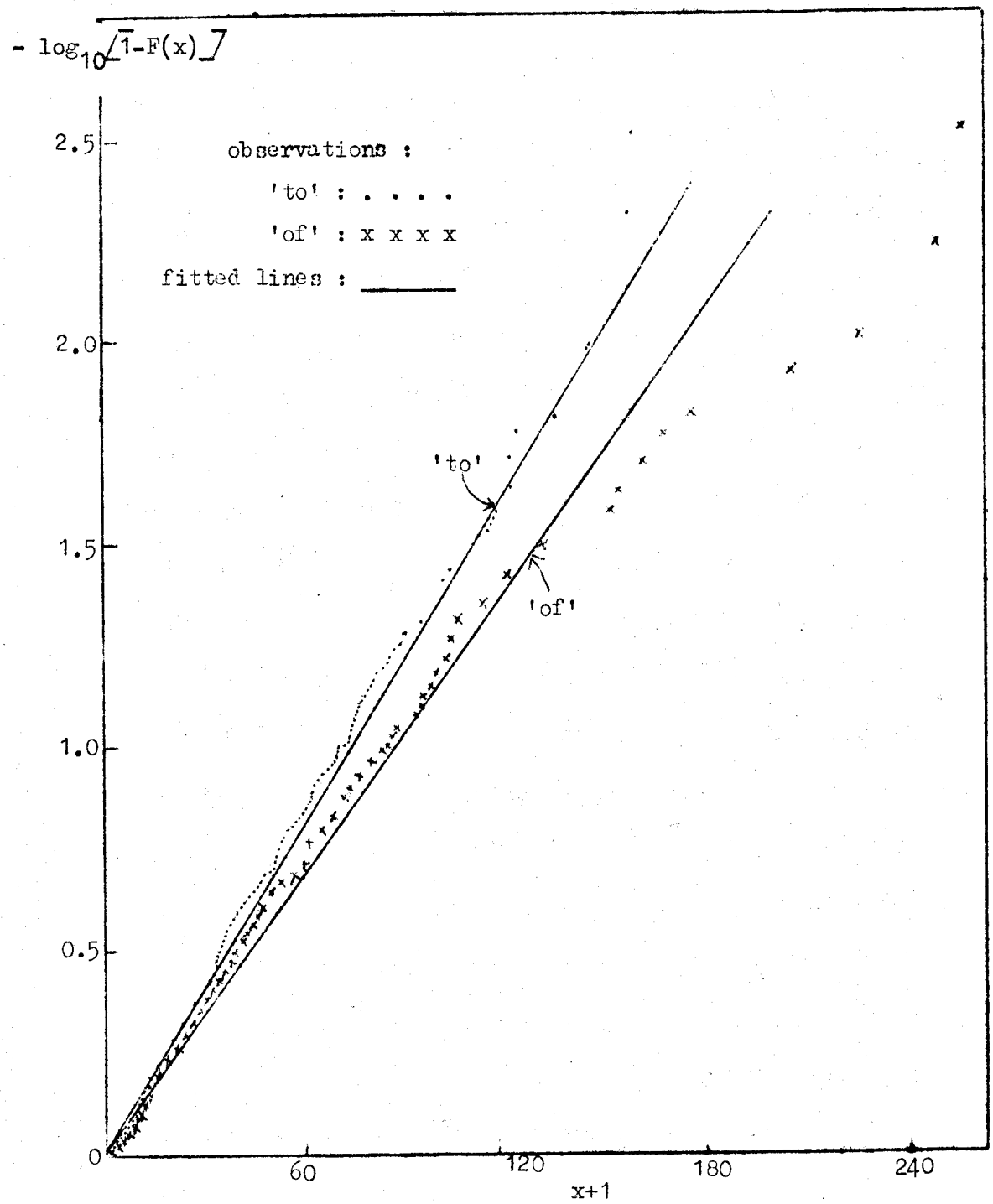


Fig. 11.4 (b): Showing $-\log_{10}[1-F(x)]$ against x+1, where x is the number of words intervening between successive occurrences of the word "to" or "of" and F(x) the cumulative proportion of intervals with length upto x: based on Chapters 1-9 of "Pride and Prejudice" by Jane Austen.

11.3.6. The Kolmogorov statistic was used to examine the goodness of fit of the geometric distribution[1]. The K-statistic for 'the' is close to the 1% level and that for 'and' to the 2.5% level; but the K for 'to' is near the 15% point and that for 'of' near the 20% point. There is no doubt that, taken together, the four K-tests show significant deviations from the geometric law. Also, the K-distances are not small, being about 6 or 8% in each case. So the geometric law gives only a rough model for the observations.

11.3.7. Intervals of length 0 are not observed in even a single case. This alone shows that the geometric distribution is a coarse approximation to reality. Very short intervals are somewhat less frequent for 'and' and 'of' than predicted by the geometric law; one occurrence tends to preclude another in the neighbourhood. For 'the', however, one finds rather too many short intervals, excepting intervals with lengths 0 and 1, and for 'to' rather too many medium-sized intervals.

11.3.8. The values of the K-statistic are much smaller for these four grammar forms than for the rare words in "Ulysses" studied in the preceding section. Clearly the geometric model is much more appropriate for these grammar forms than for the rare words occurring 5, 15 or 24 times in "Ulysses".

1/ The samples of intervals are not far from simple random (vide Section 11.5); the variate, interval, is effectively continuous; but one parameter has been estimated from the sample. On the whole, therefore, the K-test is rather conservative in the present case.

11.4.1. Some grammar forms in Bengali : A count was made of all intervals between successive occurrences of (i) the word 'se'(Bengali equivalent of 'he/she') and (ii) the word 'sei' (= the demonstrative pronoun 'that') in the whole of the Bengali novel, "Gora", by Tagore[1]. A similar count was also made, considering all words of the 'se' class (i.e., 'se', 'sei', 'sedin', 'sekhane' etc.) as a single word-type, covering roughly one-third of the same novel from the beginning. The frequency distributions are presented in Table 11.4. Lengths of intervals have been measured by the number of intervening words, as in the preceding section. Owing to lack of spaces the distributions are given in a somewhat grouped manner.

11.4.2. The graphs of log $[1 - F(x)]$ against log x are appreciably curved in all three cases, so that the zeta distribution of Zipf would fail for these grammar words of Bengali also. As before, the graphs are all concave to the horizontal axis.

11.4.3. The graphs of $-\log [1 - F(x)]$ against x are encouraging in all three cases, but the graphs are a little concave to the x-axis and show some decline in the slope over the lower ranges of x. This means, as will be seen below, that the shorter intervals are rather too frequent in comparison with the geometric model.

1/ The total number of words in "Gora" is 126,359, according to the count made by the Linguistic Research Unit of the Indian Statistical Institute. Vide Appendix 5 for the rank-frequency relation for words in "Gora".

Table 11.4: Frequency distribution of intervals according to lengths in number of intervening words between successive occurrences of the same word, separately for the words 'se'(he/she) and 'sei' (that) and for words of the 'se' class in Tagore's "Gora"*

length (no.of words)	number of intervals			length (no.of words)	number of intervals			length (no.of words)	number of intervals		
	'se'	'sei'	'se', 'sei', etc.		'se'	'sei'	'se', 'sei', etc.		'se'	'sei'	'se', 'sei', etc.
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
0	0	0	15	76-80	21	2	17	626-650	1	5	-
1	4	4	5	81-85	22	8	17	651-675	4	3	-
2	12	2	15	86-90	35	8	13	676-700	-	2	-
3	20	5	22	91-95	17	3	6	701-725	1	5	-
4	18	6	19	96-100	19	9	12	726-750	1	4	-
5	26	3	19	101-110	44	11	14	751-775	3	-	-
6	25	3	23	111-120	35	9	16	776-800	1	1	-
7	24	4	19	121-130	29	8	4	801-825	-	1	-
8	30	5	21	131-140	21	6	10	826-850	-	2	-
9	19	4	19	141-150	25	13	13	851-875	-	4	-
10	20	2	13	151-160	22	7	7	876-900	-	-	-
11	23	4	16	161-170	19	6	8	901-925	-	2	-
12	15	2	16	171-180	15	8	6	926-950	1	2	-
13	21	1	16	181-190	21	8	3	951-975	-	3	-
14	24	3	23	191-200	17	6	5	976-1000	-	-	-
15	19	2	11	201-225	36	22	7	1001-1050	-	2	
16	13		9	226-250	18	8	3	1051-1100	-	4	
17	17	2	16	251-275	16	13	4	1101-1150	-	2	
18	16	2	22	276-300	15	11	2	1151-1200	-	2	
19	13	2	10	301-325	11	12	1	1201-1250	-	1	
20	21	1	20	326-350	16	10	2	1251-1300	-	2	
21-25	71	11	54	351-375	7	10	2	1301-1350	-		
26-30	73	9	60	376-400	8	15	1	1351-1400	-	2	
31-35	51	12	30	401-425	9	10		1401-1450	-	1	
36-40	39	9	39	426-450	4	8	1	1451-1500	-	1	
41-45	36	11	32	451-475	7	2		1501-1600	-		
46-50	38	9	41	476-500	4	10		1601-1700	-	1	
51-55	38	6	25	501-525	2	4		1701-1800	1	-	
56-60	46	12	26	526-550	2	3		1801-1900	-	2	
61-65	26	8	26	551-575	1	5		1901-2000	-	1	
66-70	30	12	19	576-600	2	1		2001-2500	-	1	
71-75	29	6	14	601-625	1	4		2500-3000	-	1	
								total	1391	480	889

* Only first 200 pages were covered for the count on the 'se' class of words; for 'se' and 'sei' the whole work was covered.

11.4.4. For 'se' and 'sei', since the whole work was covered, the estimates of p were 1391/126359 = 0.01101 and 480/126359 = 0.003799, respectively. For the 'se' class of words, since $\bar{x}$ was found to be 49.3284, $\hat{p}$ came out as 0.02027. The expected distributions of intervals were calculated from the relation $\hat{F}(x) = 1 - \hat{q}^{x}$, where $\hat{q} = 1 - \hat{p}$. The observed and expected distributions are shown in Table 11.5 in the cumulative form.

11.4.5. The K-distances are of the order of 6 to 9 per cent. In this respect the fits can be regarded as fair, for practical purposes, like those found in the foregoing section. But the sample sizes are much larger for two of the three distributions here than for the four distributions in the preceding section. So, the K-statistics are more highly significant here. For 'se', the K-value is far above the 0.1% level, for 'sei', it is just below the 0.1% level, and for the 'se' class, it is a little above the 2.5% level of significance[1].

11.4.6. Detailed comparison of observed and expected frequencies shows that in general the observed frequencies are on the higher side in the lower ranges of x, excepting at the very lowest values like 0 or 1. For 'se' and the 'se' class this range extends upto about 30, but for 'sei' it reaches upto 70, broadly, although the differences are marked upto only 15.[2]

1/ For applicability of the K-test, see footnote to paragraph 11.3.6.

2/ For 'sei', it may be incidentally remarked, the lognormal distribution gave precisely the same type of fit as for words occurring 24 times in "Ulysses" (Cf. Fig.11.3).

Table 11.5: Cumulative frequency distribution of intervals according to length obtained from the frequency distributions of Table 11.4, along with the "expected" distribution based on the geometric law.

length (no.of words)*	observed and "expected" culumative percentages of intervals					
	'se' (he/she)		'sei' (that)		'se' class	
	obs.	exp.	obs.	exp.	obs.	exp.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	0.00	1.10	0.00	0.38	1.69	2.03
1	0.29	2.18	0.83	0.76	2.25	4.02
2	1.15	3.26	1.25	1.13	3.94	5.96
5	5.75	6.42	4.17	2.26	10.69	11.56
10	14.25	11.46	7.92	4.10	21.37	20.17
15	21.59	16.23	10.42	5.91	30.60	27.94
20	27.33	20.74	11.88	7.68	39.26	34.96
30	37.68	29.05	16.04	11.13	52.08	47.06
40	44.15	36.48	20.42	14.45	59.84	56.82
50	49.47	43.14	24.58	17.65	68.05	64.81
70	59.54	54.43	32.50	23.68	78.85	76.64
100	69.82	67.31	40.00	31.92	87.74	87.36
150	80.89	81.20	49.79	43.72	94.15	95.46
200	87.65	89.19	57.08	53.47	97.45	98.37
300	93.76	96.43	68.33	68.20	99.23	99.79
400	96.78	98.82	78.12	78.26	99.89	99.98
500	98.51	99.61	84.58	85.15	100.00	100.00
1000	99.93	100.00	95.21	97.79		
1500	99.93	100.00	98.75	99.67		
2001	100.00	100.00	99.58	99.95		
no. of intervals	1391		480		889	
K-distance (%)	8.63		8.81		5.02	

* Only selected values shown in the table.

11.4.7. It is clear that the geometric distribution cannot be the true probability model for the distributions in Sections 11.2 -11.4. In most cases, the deviations are significant. It is a first approximation to the unknown true model for the grammar words in English and Bengali covered in Sections 11.3 -11.4; but for the rare words in "Ulysses" (Section 11.2) the geometric law cannot be regarded as even a coarse first approximation. In all cases, interestingly enough, short intervals are rather too frequent compared with the geometric distribution.

11.5.1. Randomness of series of intervals: So far we have examined the form of the frequency distribution of the lengths of intervals pooling together all the intervals of the same word. This approach is meaningful if only the different intervals of the same word form a random series, that is, may be treated as successive observations of a random sample from the same population. We have therefore sought to demonstrate the approximate randomness of the series of intervals by applying, among other things, several non-parametric tests.

11.5.2. Table 11.6 presents the results of four types of non-parametric tests to the series of intervals of the four English grammar forms in "Pride and Prejudice" and the three Bengali grammar forms[1/] in "Gora". These tests were applied in Chapter 9 on the different series of sentence-lengths, and the description of these tests may be found there. Since the words of "Ulysses" examined in Section 11.2 occurred only 5, 15 or 24 times, none of these methods seemed to be appropriate for the short series of intervals for these words. (See below for some examination of these short series.)

1/ The third form is really a whole class of forms including 'se', 'sei' etc.

Table 11.6: Tests of randomness of the series of intervals (in terms of words) between successive occurrences of certain grammar words in English and Bengali.

name of work	word(s)	no. of intervals (n)	test based on no. of turning points (p)		no. of phases of different lengths (expected frequencies within brackets)			Wallis-Moore χ^2
			p	critical ratio	1	2	3 or more	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pride and Prejudice (Chap.1-9)	1. the	444	287	-0.808	175 (183.75)	79 (80.67)	32 (29.25)	0.710
	2. to	415	267	-0.914	157 (171.67)	81 (75.35)	28 (27.32)	1.694
	3. and	362	239	-0.062	153 (149.58)	58 (65.63)	27 (23.78)	1.401
	4. of	349	230	-0.106	148 (144.17)	56 (63.25)	25 (22.92)	1.122
Gora	5. 'se' (he/she)	1391	904	-1.545	560 (580.00)	242 (255.02)	101 (92.65)	2.107
	6. 'sei' (that)	480	332	1.392	221 (198.75)	78 (87.27)	32 (31.65)	3.479
	7. 'se', 'sei' etc.*	889	598	0.598	382 (368.33)	154 (161.88)	61 (58.78)	0.975

(contd)

name of work	word(s)	rank correlation coefficient (τ) between mean/variance of subgroups of intervals and the serial number of subgroups					
		no. of sub-groups	size of sub-groups	τ for mean	critical ratio	τ for variance	critical ratio
(1)	(2)	(10)	(11)	(12)	(13)	(14)	(15)
Pride and Prejudice (Chap.1-9)	1. the	44	10	0.023	0.212	0	0
	2. to	20 (41)	20 (10)	0.053 (0.066)	0.292 (0.595)	0.042 (0.012)	0.227 (0.101)
	3. and	40	9	0.054	0.477	0.044	0.384
	4. of	43	8	0.185	1.737	0.107	1.005
Gora	5. 'se' (he/she)	54	25	-0.103	-1.078	0.216	2.298
	6. 'sei' (that)	40	12	0.080	0.711	-0.008	-0.058
	7. 'se', 'sei', etc.*	40	22	-0.010	-0.081	0.031	0.268

* Only about 200 pages of the work were covered for this class of words

11.5.3. The turning points test does not give any significant result. For 'se', the observed number p is near the lower 5% point, but 'sei' shows nearly the same value of the critical ratio with a positive sign. Significance is not reached even by combining the tests.

11.5.4. As regards the Wallis-Moore χ^2 test, none of the χ^2's is near the significance levels, the highest being only 3.479. The seven χ^2's rather seem to be too low, averaging only about 1.6 or 1.7, $\frac{6}{7}$ths of which is much less than 2, the average of a χ^2 with 2 degrees of freedom. It is difficult to interpret this result. One may note, in this connection, that observed number of phases of length 2 is usually lower than the corresponding expected number, while the observed number of length 3 or more is always larger than the expected number.

11.5.5. The tests based on rank correlation coefficients do not reveal any significant trend in the means or variances of the subgroups. A few of the critical ratios reach the one-sided 5% level, but, on the whole, neither column of τ's is significantly different from zero. The τ's are also small in the absolute sense, mostly within $\pm$ 0.05 or $\pm$ 0.10.

11.5.6. Fractile graphical analysis also pointed to the approximate randomness of the interval series. Table 11.7 illustrates the method for the word 'sei'. Fig. 11.5(a)-(b) show the means and standard deviations of subgroups in the form of fractile graphs. It is clear, although the basis is semi-intuitive, that neither the means nor the s.d.'s show any appreciable trend.

Table 11.7: Fractile group averages and standard deviations of interval-lengths in words between 481 successive occurrences of the word 'sei' (that) in Tagore's "Gora", groups being based on the serial order of occurrence.

fractile group no.	half-sample 1			half-sample 2			combined		
	srl. no. of interval*	average length in words	s.d. in words	srl.no. of interval*	average length in words	s.d. in words	srl.no. of interval*	average length in words	s.d. in words
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(9)	(9)	(10)
1	1-12	119.08	119.91	13-24	287.08	363.49	1-24	203.08	278.26
2	25-36	87.83	75.79	37-48	243.08	206.17	25-48	165.46	171.36
3	49-60	533.58	515.57	61-72	134.42	145.22	49-72	334.00	422.82
4	73-84	307.67	492.24	85-96	375.17	455.96	73-96	341.42	465.29
5	97-108	305.08	323.58	109-120	192.83	155.29	97-120	248.96	254.79
6	121-132	241.50	252.85	133-144	296.50	346.13	121-144	269.00	297.76
7	145-156	144.50	264.59	157-168	270.67	248.13	145-168	207.58	259.00
8	169-180	355.17	371.59	181-192	407.75	536.54	169-192	381.46	452.15
9	193-204	232.83	383.81	205-216	172.33	221.69	193-216	202.58	308.08
10	217-228	184.92	181.32	229-240	411.33	342.85	217-240	298.12	292.08
11	241-252	249.67	183.02	253-264	513.42	620.06	241-264	381.54	466.95
12	265-276	218.17	287.62	277-288	377.00	584.04	265-288	297.58	457.47
13	289-300	212.17	320.51	301-312	310.25	295.82	289-312	261.21	305.77
14	313-324	329.67	222.22	325-336	95.00	89.64	313-336	212.33	204.51
15	337-348	261.67	195.55	349-360	387.33	427.14	337-360	324.50	331.16
16	361-372	321.58	260.48	373-384	134.42	197.49	361-384	228.00	245.44
17	385-396	233.33	171.99	397-408	354.17	369.50	385-408	293.75	288.54
18	409-420	270.25	342.84	421-432	275.83	233.20	409-432	273.04	286.76
19	433-444	229.25	217.24	445-456	453.08	802.84	433-456	341.17	586.43
20	457-468	355.08	328.56	469-480	118.75	122.59	457-480	236.92	270.90

* in order of occurrence in the text.

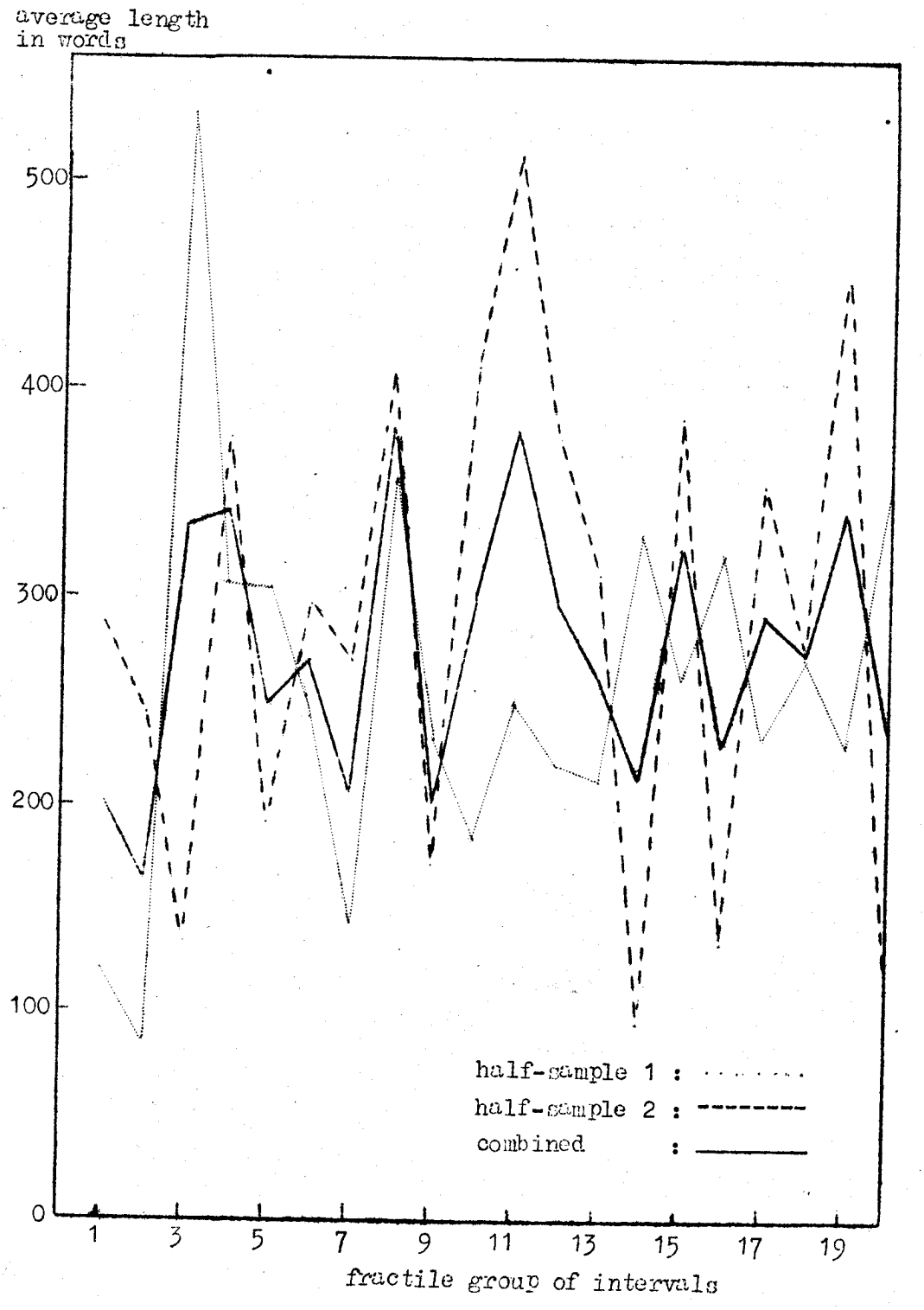


Fig.11.5 (a): Fractile group averages of interval-lengths in terms of words between 481 successive occurrences of the word 'sei'(that) in Tagore's "Gora", groups being based on the order of occurrence (vide Table 11.7).

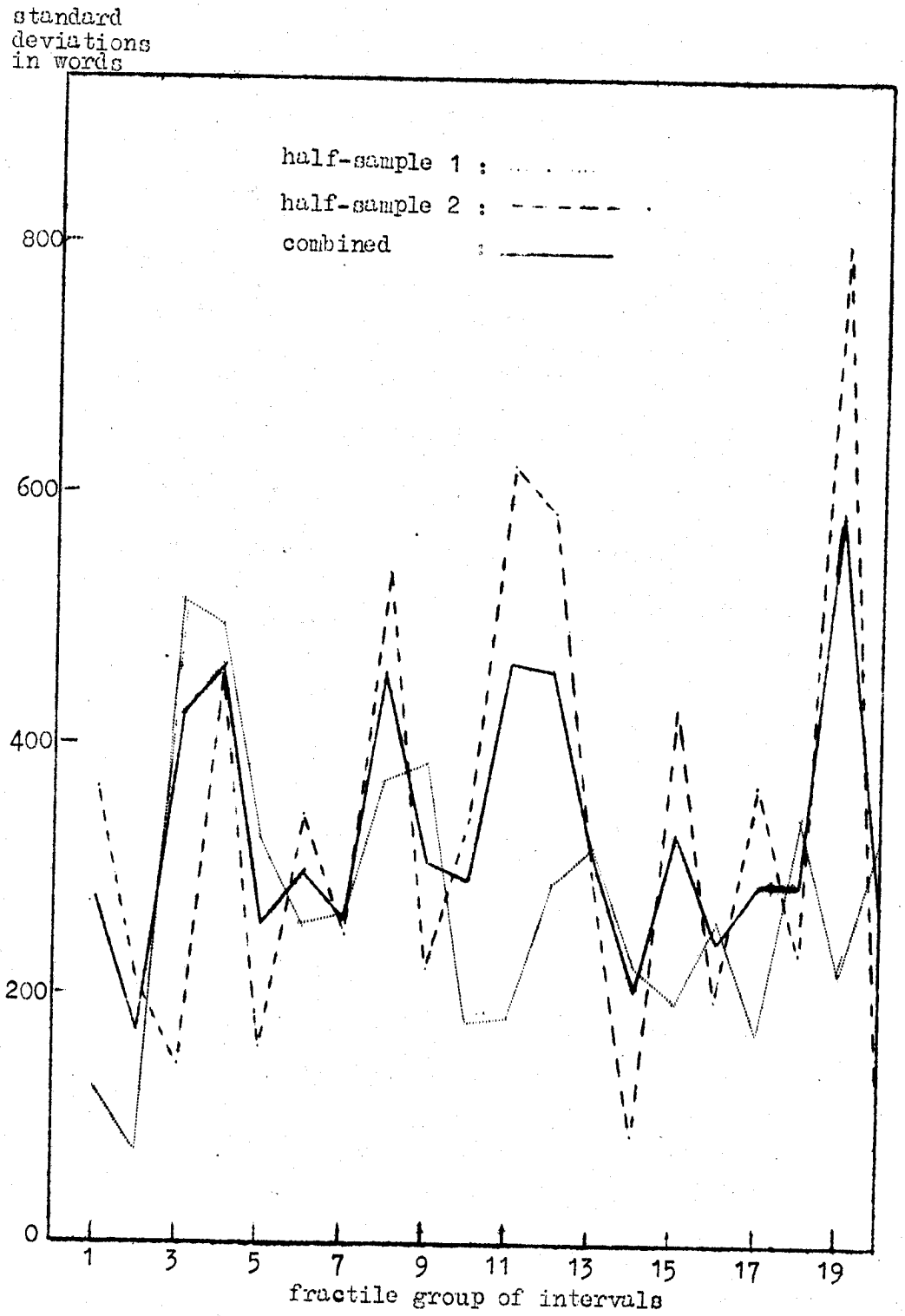


Fig.11.5 (b): Fractile group standard deviations (s.d.'s) of interval-lengths in terms of words between 481 successive occurrences of the word 'sei' (that) in Tagore's "Gora", groups being based on the order of occurrence (vide Table 11.7).

11.5.7. We now describe some tests on the Zipf material examined in Section 11.2. Each of the 898 words occurring 5 times in "Ulysses" gave four intervals (x_1, x_2, x_3 and x_4) between successive occurrences. This gave four frequency distributions of 898 intervals each, one for each of x_1, x_2, x_3 and x_4. These were compared, pair-wise, by means of the two-sample Kolmogorov statistic. The distributions, it may be noted, are effectively continuous.

11.5.8. The K-distances came out as follows:

	x_2	x_3	x_4
x_1	8.02%	8.91%	6.57%
x_2		4.23%	2.90%
x_3			4.23%

The first two values are above the 1% level and the third at 5% level; but the remaining three are well below the 20% level. Apparently, the distribution of x_1 is somewhat different from those of x_2, x_3 and x_4. This might happen, as already stated, if the topic or content of the work changes.

11.5.9. For the 97 words occurring 15 times each in "Ulysses", we calculated the correlation coefficients between certain pairs of the fourteen intervals, x_1, x_2, ..., x_{14}, taking the 97 words as the 97 observations. The correlation coefficients $r(x_i, x_j)$ were found to be as under :

$r(x_1, x_2) = 0.008$, $r(x_6, x_7) = 0.236$, $r(x_{13}, x_{14}) = 0.124$

$r(x_1, x_3) = -0.110$, $r(x_6, x_8) = -0.077$, $r(x_{12}, x_{14}) = 0.086$

$r(x_1, x_4) = -0.192$ and $r(x_{11}, x_{14}) = -0.006$

The two-sided percentage points of r are about 0.2 for the 5% level and 0.26 for the 1% level. On the whole, the correlations are small in the absolute sense and the deviations from zero are not quite significant.

11.5.10. Intervals of the 33 words each occurring 24 times in "Ulysses" were utilised for calculating $r(x_i, x_{i+1})$ for all values of i(viz., 1, 2, ..., 22). Here, we denote the 23 intervals of a word by $x_1, x_2 \ldots, x_{23}$. The correlation coefficients turned out to be[1/]:

$r(1,2) = 0.143$, $r(2,3) = -0.197$, $r(3,4) = -0.129$, $r(4,5) = 0.092$

$r(5,6) = 0.335$, $r(6,7) = -0.004$, $r(7,8) = -0.018$, $r(8,9) = -0.099$

$r(9,10) = -0.011$, $r(10,11) = 0.026$, $r(11,12) = 0.198$, $r(12,13) = -0.059$,

$r(13,14) = -0.028$, $r(14,15) = -0.139$, $r(15,16) = -0.145$, $r(16,17) = 0.175$,

$r(17,18) = 0.123$, $r(18,19) = 0.193$, $r(19,20) = 0.326$, $r(20,21) = 0.083$,

$r(21,22) = -0.074$ and $r(22,23) = 0.028$.

The two-sided 5% and 1% points are about 0.345 and 0.443 respectively. Only two of the 22 values are near + 0.345. The signs and magnitudes of the correlation coefficients show that they may easily be regarded as a random sample from the wellknown sampling distribution of r for 31 degrees of freedom. The sequence of r's also do not present any systematic feature. To all appearances, the r's are not significantly different from zero or indicative of any type of deviation from randomness.

1/ $r(x_i, x_j)$ written here as r(i,j).

Appendix 1 : Word-length in letters in Bengali prose

1. Word-length has been measured in syllables for the studies on Bengali prose and poetry reported in the main text. Some statistics are presented here for word-length in letters in two representative novels in Bengali prose, viz., "Visavriksha" by Bamkimchandra and "Sheser Kavita" by Tagore.

2. In each case, only the probability sample of words falling on 100 randomly selected lines was utilised for this purpose. The sample comprised four independent and interpenetrating subsamples, each covering 25 sample lines (vide Chapter 2, Sections 2.3, 2.5). Composition of each sample word in terms of letters was determined following standard Bengali orthography. Authors' spelling was accepted. The following conventions had to be adopted:

(a) The silent অ ('a') appearing after consonants (and hence not written) in medial or terminal positions was counted as a letter provided the word belonged to the 'tatsama' (Sanskrit) group of words.

(b) The double consonants like ব্ব ('bb') or ত্ত ('tt') were counted as two letters.

(c) The 'chandra-vindu' ঁ (n) was counted as one letter.

(d) The 'mahaprana' (aspirate) letters like খ (kh), ঘ (gh) were taken as single letters.

(e) The two b's (ব ব) were not distinguished and the whole frequency was recorded against the very much more frequent 'ব' of 'p'- group (প-varga).

3. Table 1 shows the word-length distributions, averages etc., in terms of letters.

Table 1 : Word-length in letters in "Visavriksha" and "Sheser Kavita" separately by subsamples, based on probability samples of words falling on 100 randomly selected lines.

no. of letters	percentage of sample words									
	Visavriksha					Sheser Kavita				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	0.66	2.06	-	1.80	1.15	0.53	0.57	0.54	-	0.41
2	7.28	4.79	5.44	9.58	6.87	7.98	5.06	5.95	10.33	7.35
3	5.96	6.16	7.48	6.59	6.55	10.64	7.87	9.19	11.96	9.93
4	19.20	16.45	21.78	13.77	17.68	26.06	28.09	32.97	29.34	29.12
5	13.91	22.61	17.01	17.36	17.68	25.54	21.35	18.92	16.85	20.68
6	26.49	18.49	15.65	15.57	18.98	11.17	13.48	15.68	13.04	13.33
7	9.28	4.79	6.80	14.37	9.00	5.32	6.18	7.03	8.70	6.80
8	5.30	8.22	6.80	8.98	7.36	5.32	7.30	3.78	5.44	5.44
9	3.97	5.48	5.44	2.99	4.42	3.72	3.93	3.24	1.63	3.13
10	4.64	4.11	6.12	4.79	4.91	3.19	2.81	1.08	0.54	1.90
11	2.65	1.37	3.40	1.80	2.29	-	1.12	-	0.54	0.41
12	-	1.37	2.04	1.20	1.15	-	0.56	1.08	-	0.41
13	0.66	-	2.04	1.20	0.98	0.53	1.12	0.54	-	0.54
14	-	2.74	-	-	0.66	-	0.56	-	1.09	0.41
15	-	0.68	-	-	0.16	-	-	-	-	-
16	-	0.68	-	-	0.16	-	-	-	-	-
23	-	-	-	-	-	-	-	-	0.54	0.14
total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
average (letters)	5.656	6.007	6.000	5.731	5.843	4.995	5.410	5.000	4.946	5.084
s. d. (letters)	2.246	2.841	2.642	2.495	2.566	2.219	2.280	1.940	2.428	2.271
average (syllables)	2.404	2.534	2.531	2.419	2.470	2.122	2.371	2.173	2.125	2.186
s. d. (syllables)	0.907	1.166	1.114	1.023	1.056	0.839	0.843	0.865	1.064	0.938
no. of sample words	151	146	147	167	611	188	178	185	184	735

4. The mode for the distribution is at 4 letters for "Sheser Kavita", but for "Visavriksha" three lengths — 4,5 and 6 letters — have very nearly similar frequencies. This may be due to the greater difference in length between conversational and other words in "Visavriksha"[1]. The median length is again more or less definitely 5 letters for "Sheser Kavita", but between 5 and 6 letters for "Visavriksha".

5. The close correlation between word-length in syllables and word-length in letters will be demonstrated below. Table 1 shows that the average word-length in letters is about 2.35 times the average word-length in syllables. This ratio shows a very high degree of consistency. The ratio between the s.d.'s is, however, somewhat erratic for "Sheser Kavita". On the whole, the s.d. in letters is about 2.4 times the s.d. in terms of syllables. The coefficient of variation of word-length is about 44% in terms of letters and 43% in terms of syllables.

6. In view of the close correlation between word-length in letters and word-length in syllables, the estimates of average length in syllables given in Chapter 3 can lead to fairly good estimates of average length in letters. "Visavriksha" and "Sheser Kavita" represent two extremes of modern Bengali prose; so the average word-length in most Bengali works should range between 5 and 6 letters. For most modern works, the average should be less than 5.5 letters.

1/ "Visavriksha" has a small percentage of conversational words according to the strict definition of conversational words adopted in Chapter 4. But if soliloquies, letters etc., are included in conversations, the percentage would rise appreciably.

7. Figure 1 is presented for showing how far the observed word-length distributions of Table 1 can be regarded as grouped versions of underlying lognormal distributions. (Vide Chapter 6 for details of fitting lognormal distribution to distributions of word-length in syllables.) In one approach, the observed values 1, 2, 3, ... are assumed to represent the intervals 0-1, 1-2, 2-3, ... respectively of the underlying continuous variate; in the second approach, the intervals are assumed to be 0-1.5, 1.5-2.5, 2.5-3.5, ... As in several places in the main text, the four subsamples from each work were merged into two half-samples for graphical purposes. The graphs show that the fit should be generally satisfactory, especially by the second approach.

8. Finally, we present in Table 2 the joint distribution of the number of syllables and the number of letters comprising each sample word. Only the combined distribution is presented for each work, but the correlation coefficients are given in Table 3 separately by subsamples as well as for the combined sample.

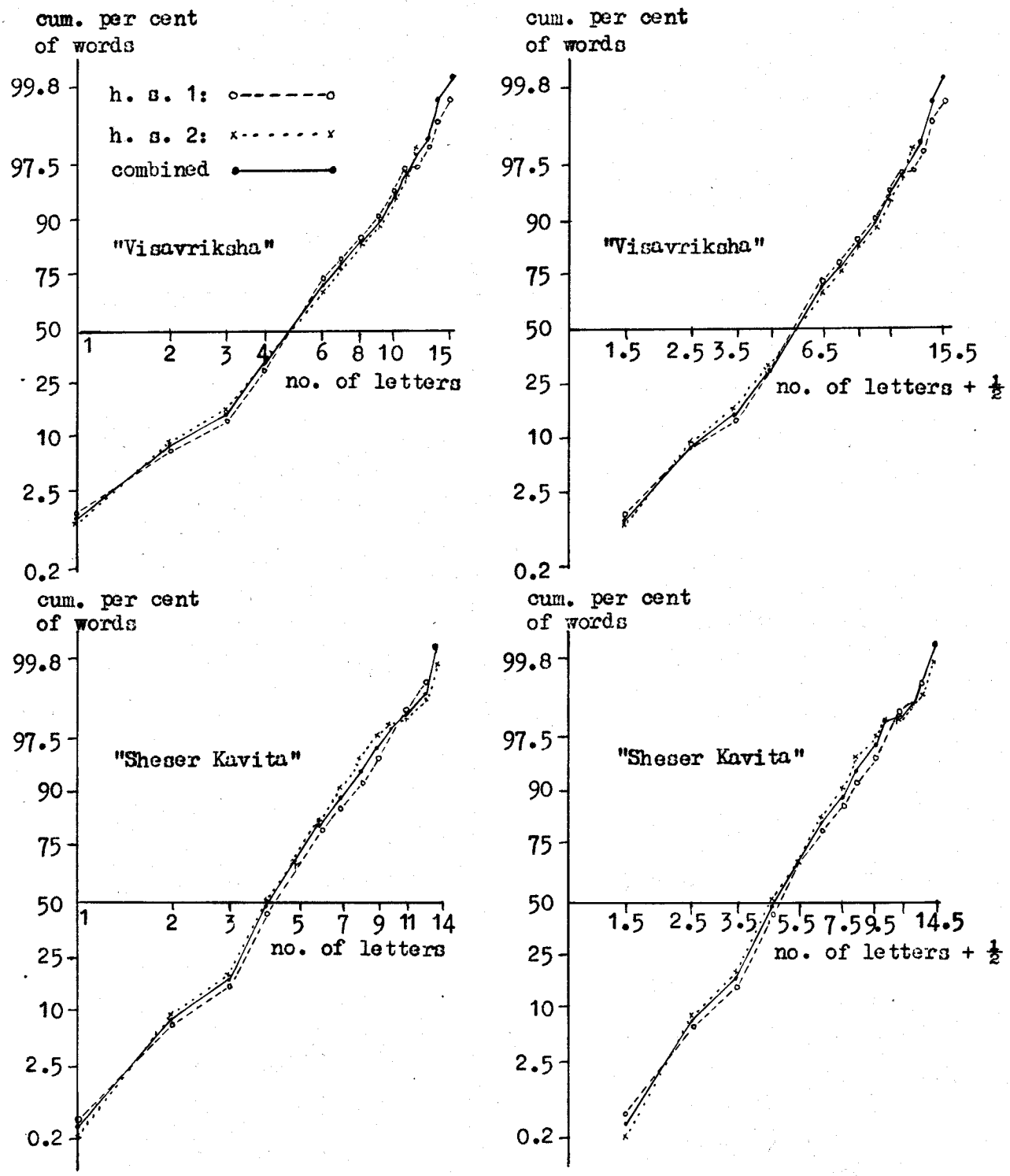


Fig. 1: Ogives, on log-probit scale, for distributions of word-length in letters (x) estimated from probability samples of 611 and 735 words from "Visavriksha" and "Sheser Kavita" respectively, assuming that x = 1, 2, 3, ..., represent (i) intervals 0-1, 1-2, 2-3, ..., and (ii) intervals 0-1.5, 1.5-2.5, 2.5-3.5, ..., of the underlying lognormal variate.

Table 2: Joint distribution of words by length in syllables and length in letters, based on probability sample of words falling on 100 randomly selected lines

no. of letters	no. of syllables								total
	1	2	3	4	5	6	7	11	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				(a) Visavriksha					
1	7								7
2	42								42
3	30	10							40
4	15	93							108
5	5	94	9						108
6		38	78						116
7		11	42	2					55
8		2	31	12					45
9			10	17					27
10			5	21	4				30
11			1	8	5				14
12				2	3	2			7
13					5	1			6
14					1	1	2		4
15					1				1
16					1				1
total	99	248	176	62	20	4	2	-	611
				(b) Sheser Kavita					
1	3								3
2	54								54
3	60	13							73
4	22	192							214
5	1	141	10						152
6		47	51						98
7		6	42	2					50
8			24	16					40
9			10	12	1				23
10			2	8	4				14
11				2	1				3
12					3				3
13					3	1			4
14					3				3
23								1	1
total	140	399	139	40	15	1	-	1	735

9. The Tables 2(a) and 2(b) point to the close correlation between number of syllables and number of letters comprising a word. The correlation coefficients (r) are given below. The regressions are sensibly linear.

Table 3. Product-moment correlation coefficient (r) between number of syllables and number of letters comprising sample words

work	r by subsamples of probability sample*				
	SS 1	SS 2	SS 3	SS 4	combined
(1)	(2)	(3)	(4)	(5)	(6)
Visavriksha	0.900 (151)	0.909 (146)	0.907 (147)	0.903 (167)	0.905 (611)
Sheser Kavita	0.818 (188)	0.917 (178)	0.892 (185)	0.924 (184)	0.873 (735)

* No. of sample words given inside brackets

10. It is felt that the findings on word-length would have been more or less unaffected if word-length had been measured in letters instead of syllables. Since the number of letters would be closely correlated with the number of phonemes so far as Bengali is concerned, the same may be said about the use of phonemes also.

....

Appendix 2 : Concentration Curve of Word-length and Sentence-length

1.1. Concentration Curves for Discrete Size Distributions : The Gini-Lorenz concentration curve (vide Kendall and Stuart, 1958, Vol.1, pp. 48-51) is usually defined for size distributions of continuous variates like income. It has been shown (Bhattacharya and Mahalanobis, 1964) that this restriction is not really necessary, that is, the definition can be extended to size distributions of discrete variates. The variate must, of course, be non-negative and must be such that the idea of the shares of different groups of individuals is meaningful. It will be assumed below that the variate x assumes the values 0,1,2,....., but this assumption is really unnecessary.

1.2. Denote by p_r the probability that x assumes the value r ($r = 0,1,2,.....$). Then $\sum_r p_r = 1$. Assume that the mean $\sum_x xp_x$ exists and denote this by μ. Then

$$\Phi_r = \sum_0^r xp_x / \mu \quad (r = 0,1,2,.....)$$

can be interpreted as the proportion of the total amount of the variate x enjoyed or covered by those individuals of the population whose x-values are less than or equal to r. Plotting Φ_r against the cumulative proportion

$$F_r = \sum_0^r p_x \qquad (r = 0,1,2,.....)$$

one gets a series of points including the point (1,1), which forms the basis of the concentration curve. The point (0,0) can be added to these.

1.3. The problem arises as to whether one can join these points in any meaningful way and get a continuous concentration curve analogous to that for continuous variates.

1.4. If the individuals having the same value of x are ranked arbitrarily, that is to say, if the randomization method is adopted for getting a ranking by x-values without ties, then the concentration curve would be the broken chain obtained by joining the above-mentioned points successively by straight lines.

1.5. This kind of joining has a stronger support. One may recall that the Gini-Lorenz concentration coefficient can be defined in two ways : first, as twice the 'area of concentration' between the egalitarian line and the continuous concentration curve, and second, through the Gini mean difference, a measure of variability. In the continuous case, both the approaches lead to the same measure. In the discrete case, the first approach cannot be made at all, for the concentration curve is to start with only a series of discrete points, but the concentration coefficient can be meaningfully defined in terms of the Gini mean difference. Now it can be shown that, in the discrete case, if the continuous concentration curve is drawn as described in para 1.4, twice the area of concentration obtained agrees with the definition of the concentration coefficient in terms of the Gini mean difference.

1.6. For, we get for the Gini mean difference (Δ_1) of the discrete x-distribution

$$\Delta_1 = E \left| x_1 - x_2 \right|$$

(where x_1 and x_2 are two independent observations from the population under study)

$$= 2 \sum_{x_2 \leq x_1} (x_1 - x_2)\, p_{x_1}\, p_{x_2}$$

$$= 2 \sum_{x_1} \left[x_1\, p_{x_1}\, F_{x_1} - p_{x_1}\, \mu\, \Phi_{x_1} \right]$$

$$= 2\mu \sum_{x} \left[F_x \Delta \Phi_{x-1} - \Phi_x \Delta F_{x-1} \right]$$

(where Δ denotes the first difference). Geometrical considerations show that the concentration coefficient defined as $\frac{\Delta_1}{2\mu}$, that is, as $\sum_x \left[F_x \Delta \Phi_{x-1} - \Phi_x \Delta F_{x-1} \right]$, exactly equals twice the area between the egalitarian line $\Phi = F$ and the broken chain concentration curve defined in para 1.4.

1.7. For computational purposes the concentration coefficient may be expressed as

$$G = \frac{\Delta_1}{2\mu} = 1 - \sum_x p_x \left(\Phi_x + \Phi_{x-1} \right)$$

1.8. It may be of interest to study the concentration curve for the Poisson distribution, which according to Fucks (1955) fits the distribution of word-length in syllables. (Chapter 6 shows that

this is far from true in Bengali prose.) Let λ denote the Poisson parameter. Then $\mu = E(x) = \lambda$ and

$$F_r = e^{-\lambda} \sum_{0}^{r} \frac{\lambda^x}{x!}$$

$$\text{and } \Phi_r = \sum_{0}^{r} e^{-\lambda} \frac{\lambda^x}{x!} \; x/\lambda = e^{-\lambda} \sum_{0}^{r} \frac{\lambda^{x-1}}{(x-1)!} = F_{r-1} \Bigg\} (r=0,1,2,\ldots)$$

Also $F_{-1} = 0$. Hence $F_r - \Phi_r = p_r$. Fucks, however, considered x-1 (where x = word-length in syllables) as the Poisson variate and not x.

2.1. Distributions of syllables by word-length : It has been seen in Chapter 3 that the average word-length in Bengali varies from nearly 2 syllables in works like "Char-Yari Katha" to about 2.7 syllables per word in works like "Shakuntala"; a few poems show still higher averages. This is certainly not a small range. But the common reader seems to be extremely sensitive to much smaller differences in the average, and can easily detect differences of 0.2 syllables per word or even less. Apparently, the reader is particularly sensitive to the relative frequencies of longer words, with (say) four syllables or more; and these relative frequencies increase more than proportionately when the average word-length rises. Thus, the proportion of words with more than three syllables is about 12 or 13 % for "Ghare Baire", about 28% for "Durgeshnandini" and about 33% for "Shakuntala"; the corresponding $\bar{x}$-values are 2.09, 2.58 and 2.70 respectively. But there seems to be a subtler explanation of the sensitiveness of readers.

2.2. When a reader looks at random across the pages of a work his eyes seem to select words with probabilities roughly proportional to their lengths. If the probabilities are assumed to be exactly proportional to word lengths in syllables, the lengths of words selected by the reader form a distribution different from the distribution obtained if words are selected with equal probabilities. The former is, in fact, the distribution of randomly selected syllables according to lengths of words containing them. If $x = 1,2,3,\ldots\ldots$ denotes word-length in syllables and p_x denotes the proportion of words having length x, then the distribution mentioned above will have the relative frequency

$$q_x = \frac{xp_x}{\bar{x}} \qquad (x = 1,2,3,\ldots\ldots)$$

Such distributions have been called the first moment distributions in the literature [vide Aitchison and Brown, 1957, pp. 12-13; the literature discusses the continuous case explicitly[1]. The probability $q_4 + q_5 + \ldots$ ad inf. is much larger than the probability $p_4+p_5+\ldots$.

2.3. The distributions of syllables by word-length in syllables are presented in Table 1 for a few selected works. For the sake of interest, the corresponding distributions of words are also shown alongwith. The material used has been presented in Table 3.2 of Chapter 3. The four subsamples of words have been grouped to form two half-samples, also corresponding subsamples of probability and systematic samples were pooled where both happened to be available.

1/ An analogy may be drawn with the distribution of households by household size and the distribution of persons by sizes of respective households.

Table 1. Percentage distribution of words and syllables by word-length in syllables, for four selected works in Bengali prose.*

(a) percentage distribution of words

work	type of sample	half sample	no. of sample words	percentage of words by word-length in syllables							
				1	2	3	4	5	6	7	8
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Shakuntala	prob.	1	351	13.39	34.48	33.05	13.39	3.99	1.70	-	-
		2	345	9.86	34.78	34.49	13.91	4.93	1.74	0.29	-
		comb.	696	11.64	34.64	33.76	13.65	4.45	1.72	0.14	-
Durgeshnandini	prob.	1	1190	13.03	37.48	32.77	12.69	2.94	0.76	0.17	0.16
	plus	2	1169	15.06	36.69	32.07	9.75	4.45	1.03	0.86	0.09
	syst.	comb.	2359	14.03	37.09	32.43	11.23	3.69	0.89	0.51	0.13
Gora	-do-	1	1365	15.16	46.75	28.64	6.89	2.27	0.29	-	-
		2	1348	16.32	48.22	25.00	7.72	2.45	0.15	0.07	0.07
		comb.	2713	15.74	47.48	26.83	7.30	2.36	0.22	0.04	0.03
Ghare Baire	prob.	1	984	21.65	57.01	15.86	4.37	0.81	0.20	0.10	-
		2	917	19.63	57.47	16.03	5.23	1.42	0.22	-	-
		comb.	1901	20.67	57.24	15.94	4.79	1.10	0.21	0.05	-

* based on material presented in Table 3.2 of Chapter 3.

Table 1: (Contd.)

(b) percentage distribution of syllables

work	type of sample	half sample	no.of sample syllables	percentage of syllables by word-length in syllables 1	2	3	4	5	6	7	8
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Shakuntala	prob.	1	931	5.05	25.99	37.38	20.19	7.52	3.87	-	-
		2	951	3.58	25.23	37.53	20.19	8.94	3.79	0.74	-
		comb.	1882	4.30	25.61	37.46	20.19	8.24	3.83	0.37	-
Durgeshnandini	prob.	1	3080	5.03	28.96	37.99	19.62	5.68	1.75	0.45	0.52
	plus	2	3025	5.82	28.36	37.19	15.07	8.60	2.38	2.31	0.27
	syst.	comb.	6105	5.42	28.67	37.59	17.36	7.13	2.06	1.38	0.39
Gora	-do-	1	3211	6.45	39.73	36.53	11.71	4.83	0.75	-	-
		2	3139	7.01	41.42	32.21	13.25	5.26	0.38	0.22	0.25
		comb.	6350	6.72	40.57	34.39	12.47	5.04	0.57	0.11	0.13
Ghare Baire	prob.	1	2034	10.47	55.16	23.01	8.46	1.97	0.59	0.34	-
		2	1944	9.26	54.22	22.68	9.88	3.34	0.62	-	-
		comb.	3978	9.88	54.70	22.85	9.15	2.64	0.60	0.18	-

* based on material presented in Table 3.2 of Chapter 3.

2.4. It is but one step from this to the concentration curves in Fig. 1 showing cumulative proportion of words along the horizontal axis and the cumulative proportion of syllables along the other. The curves are given by halfsamples, but for only one of the four works covered in Table 1. The curves for different works are fairly close.

2.5. The same type of analysis can be carried out on the sentence-length distributions of Chapter 8. This case is, however, of less interest.

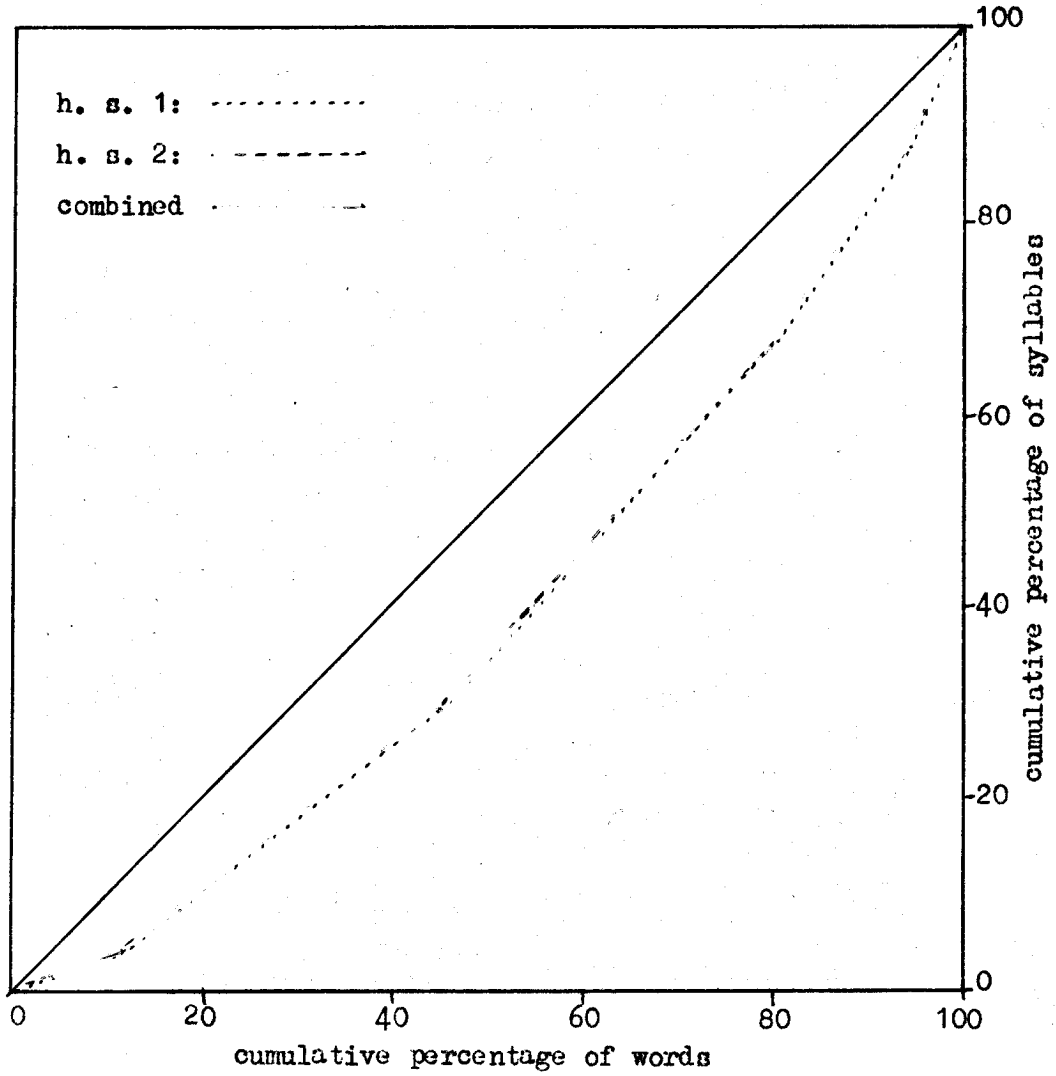


Fig.1: Concentration curves of word-length in syllables based on the probability sample of 696 words from "Shakuntala" [Vide Table 1].

Appendix 3: On word-length in Jane Austen's "Pride and Prejudice"

1. This appendix reports on some studies on word-length in the English novel, "Pride and Prejudice" by Jane Austen. The first 32 chapters (pp. 1-199) of the book were covered for the purpose[1]. A random sample of 200 lines was selected by the method described in Section 2.3 of Chapter 2; this gave the probability sample of 1772 words. A "systematic" sample of 199 lines was also chosen, by taking the third line from top of each of pages 1-199; this gave the systematic sample of 1768 words. Each sample was split into four interpenetrating subsamples in the same way as described in Chapter 2.

2. The following paragraphs discuss among other things (i) the distribution of word-length estimated for "Pride and Prejudice", (ii) the agreement between the two types of samples, (iii) the relative frequencies and lengths of conversational and other words, (iv) the homogeneity of different parts of the text covered and (v) the log-normality of the word-length distribution[2]. In general, the methodology is similar to those followed for the corresponding studies on Bengali reported in the main text : only here word-length is measured in syllables.

3. Table 1 presents the word-length distributions estimated for "Pride and Prejudice" (Chapters 1-32). The first thing to note is the close agreement between the two methods of sampling. Homogeneity χ^2

1/ This limitation of coverage has no particular reason.

2/ The study on randomness of the word-length series has been reported in Chapter 5.

for comparing the combined distributions [columns (6) and (11)] comes out as 13.262 (12 d.f.'s) which is quite reasonable. The two-sample Kolmogorov-Smirnov test was also applied. The maximum distance between the two distribution functions is 2.31% at 2 letters. This is small in an absolute sense and not significant even at the 20% level. Since both samples are more reliable than unrestricted random samples of the same size (vide next para) the above tests are somewhat conservative in nature.

4. In view of the negative autocorrelations r_1 of the word-length series (vide Section 5.6, Chapter 5) the probability sample might have smaller sampling errors than unrestricted random samples of equal size (no. of words). This is indicated by the homogeneity χ^2 for comparing the subsample distributions [columns (2) - (5) of Table 1]. This χ^2 turns out to be 25.22, which is rather low, though not significantly so, since the degrees of freedom number 33. The corresponding χ^2 for comparing the subsamples of the systematic sample [columns (7) - (10) of Table 1] comes out as 24.47 (33 d.f.). Since the two samples have nearly equal sizes, this shows that the distributions from systematic samples are just about as reliable as those from probability samples.

5. Tables 2 and 3 present some data on words used in conversations and elsewhere. Short letters, it should be noted, were included in "conversations".

Table 1. Distribution of words by length in letters estimated for Jane Austen's "Pride and Prejudice" (Chapters 1 to 32).

word-length in letters	estimated percentage of words										
	probability sample					systematic sample					pooled
	s.s.1	s.s.2	s.s.3	s.s.4	comb.	s.s.1	s.s.2	s.s.3	s.s.4	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	3.70	3.64	1.82	2.61	2.93	3.18	4.61	3.10	3.39	3.56	3.25
2	18.26	16.50	17.27	22.61	18.74	20.45	16.82	24.12	20.14	20.42	19.58
3	23.27	23.06	25.90	21.74	23.47	22.96	22.80	23.89	23.08	23.18	23.32
4	20.43	16.26	17.73	18.04	18.17	16.14	19.59	15.71	15.84	16.80	17.48
5	8.48	8.50	10.23	8.48	8.92	10.00	10.37	8.85	9.28	9.62	9.27
6	7.39	8.01	7.05	7.17	7.39	5.91	6.22	7.08	7.01	6.56	6.98
7	5.65	7.77	6.59	7.39	6.83	7.72	8.76	5.75	7.92	7.52	7.17
8	4.78	5.83	3.86	4.57	4.74	3.64	3.92	3.76	4.75	4.02	4.38
9	4.78	5.34	5.91	3.70	4.91	4.32	3.00	4.20	4.07	3.90	4.41
10	1.09	2.91	1.82	1.74	1.86	2.73	1.15	1.11	2.49	1.87	1.86
11	0.65	0.97	0.68	0.43	0.68	1.59	1.38	1.55	0.68	1.30	0.99
12	0.87	0.97	0.91	1.30	1.02	0.91	0.69	0.44	0.90	0.74	0.88
13	0.43	0.24	0.23		0.23	0.45	0.69	0.44	0.45	0.51	0.37
14	0.22	-	-	0.22	0.11					-	0.06
total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
no. of sample words	460	412	440	460	1772	440	434	452	442	1768	3540
average	4.365	4.650	4.482	4.330	4.452	4.466	4.378	4.199	4.439	4.369	4.410

Table 2: Percentage of all words used in conversations and average lengths of conversational and other words, estimated for "Pride and Prejudice",(Chapters 1 to 32).

	probability sample					systematic sample					pooled
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. percent of conv. words	44.1	48.5	35.4	58.7	46.8	39.1	47.9	45.8	39.6	43.1	44.9
2. average no. of letters in											
(a)conv. words	4.236	4.420	4.103	4.007	4.181	3.977	4.029	3.787	4.263	4.005	4.097
(b)other words	4.467	4.868	4.690	4.790	4.689	4.780	4.699	4.547	4.554	4.645	4.666
(c)all words	4.365	4.650	4.482	4.330	4.452	4.466	4.378	4.199	4.439	4.369	4.410
3. no. in sample											
(a)conv. words	203	200	156	270	829	172	208	207	175	762	1591
(b)other words	257	212	284	190	943	268	226	245	267	1006	1949
(c)all words	460	412	440	460	1772	440	434	452	442	1768	3540

Table 3: Length distributions of conversational and other words estimated for "Pride and Prejudice", (Chapters 1 to 32).

word-length in letters	percentage of words					
	conversational words			other words		
	prob. sample	syst. sample	pooled	prob. sample	syst. sample	pooled
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	4.10	6.30	5.15	1.91	1.49	1.69
2	21.23	24.01	22.56	16.54	17.69	17.13
3	23.53	23.23	23.39	23.44	23.17	23.30
4	19.18	14.83	17.10	17.29	18.29	17.81
5	8.08	10.63	9.30	9.65	8.85	9.24
6	7.36	6.04	6.73	7.42	6.96	7.18
7	5.55	5.25	5.41	7.95	9.24	8.62
8	4.10	3.15	3.65	5.30	4.67	4.97
9	3.50	2.76	3.15	6.15	4.77	5.44
10	1.69	1.97	1.82	2.01	1.79	1.90
11	0.72	1.05	0.88	0.64	1.49	1.08
12	0.84	0.39	0.62	1.17	0.99	1.08
13	0.12	0.39	0.25	0.32	0.60	0.46
14	-	-	-	0.21	-	0.10
total	100.00	100.00	100.00	100.00	100.00	100.00
no.of sample words	829	762	1591	943	1006	1949

6. The agreement between the two types of samples is further evident from Tables 2 and 3. Both types indicate that nearly 45% of all words were used in conversations. Both also point to be appreciable and obviously significant difference between lengths of the two classes of words. The average difference is around 0.55 or 0.60 letters per word. A smaller proportion of other words have only 1 or 2 letters and a larger proportion seven letters or more.

7. The estimates of standard deviation show remarkable consistency between the two types of samples. The estimates for combined samples are shown below :

	estimated s.d. of word-length in terms of letters	
	probability sample	systematic sample
(1)	(2)	(3)
1. conversational words	2.330	2.359
2. other words	2.518	2.508
3. all words	2.445	2.466

This means that coefficient of variation is about 55% for all words, 54% for other words and about 57 or 58% for conversational words.

8. We now use the systematic sample of words to apply a test of homogeneity. The sample lines from pages 1-50 gave a sample of words from what may be called "Part 1" of the work; the sample lines from pages 51-100 gave a sample from "Part 2"; and so on. Table 4 presents the word-length distributions estimated for the four different "Parts".

Table 4: Word-length distributions for four "Parts" of "Pride and Prejudice", estimated from a systematic sample of 1768 words.

word-length in letters	percentage of words			
	Part 1 (pp.1-50)	Part 2 (pp.51-100)	Part 3 (pp.101-150)	Part 4 (pp.151-199)
(1)	(2)	(3)	(4)	(5)
1	5.02	2.48	3.47	3.29
2	19.42	21.85	20.61	19.76
3	23.98	22.97	23.65	22.13
4	17.36	16.44	16.70	16.71
5	10.28	8.33	9.98	9.88
6	5.25	7.43	6.29	7.29
7	7.76	8.33	6.51	7.53
8	3.65	4.50	2.82	5.18
9	3.88	2.03	5.86	3.76
10	1.60	2.48	1.08	2.35
11	0.68	1.13	1.95	1.41
12	0.91	1.13	0.86	-
13	0.23	0.90	0.22	0.70
total	100.00	100.00	100.00	100.00
average	4.240	4.421	4.364	4.454
s. d.	2.378	2.530	2.484	2.460
no. of sample words	438	444	461	425

9. The two-sample Kolmogorov-Smirnov test was applied for comparing each pair of "Parts". The K-S distances are as follows:

Part 1 vs Part 2 - 3.99% Part 2 vs Part 3 - 2.34%

-do- vs Part 3 - 2.67% -do- vs Part 4 - 2.12%

-do- vs Part 4 - 4.29% Part 3 vs Part 4 - 2.64%

None of these reaches even the upper 20% point of the distribution which is over 7 %. Thus, inspite of the conservative nature of

the test[1], we can definitely conclude that the "Part" distributions show satisfactory homogeneity.

10. Figure 1 represents the frequently used graphical test of lognormality of the word-length distribution for all classes of words. The probability sample and the systematic sample have been used as halfsamples of the pooled sample, so as to obtain some idea about the significance of the nonlinearities. Clearly, the ogive is significantly curved, on log-probit scale[2]. The curvature is reduced, but not eliminated, if the cumulative percentages are plotted against log 1.5, log 2.5, ... instead of against log 1, log 2, ... respectively (vide Chapter 6, Section 6.5). This diagram is not shown, however. Also, the curvature appears to be less when only conversational or only non-conversational words are studied instead of all classes of words. It is possible that either class of words obeys lognormality to a closer approximation than does the population of all words (vide Section 6.7 of Chapter 6).

1/ The main reason is the departure from unrestricted random sampling (vide paragraph 3 supra). Another minor reason is the discrete nature of the word-length variate.

2/ Williams (1956) also found such non-linearities when examining Mendenhall's data on words in Shakespeare and in "Vanity Fair".

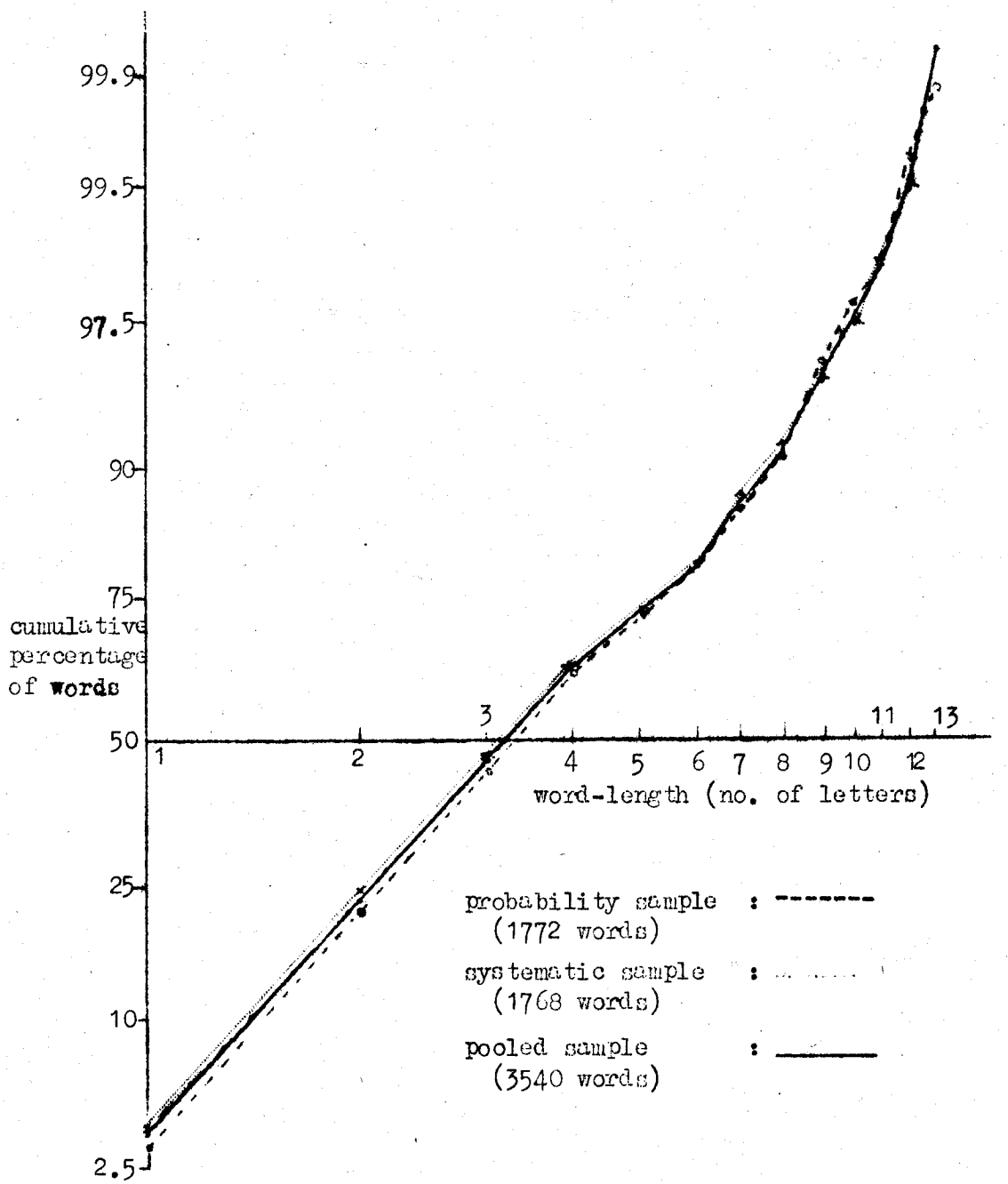


Fig. 1: Ogive, on log-probit scale, for the distribution of words by length in letters estimated for "Pride and Prejudice" (Chapters 1-32) from probability and systematic samples of words.

Appendix 4 : Relative frequencies of letters in Bengali prose

1. Statistics are now available for many languages including English, on the relative frequencies of letters, phonemes, and syllables, and also of pairs and triplets of letters or phonemes, called digrams and trigrams; relative frequencies of letters etc. in different positions of the word may also be mentioned in this connection. The uses of such data are many and varied. They are useful for pedagogic purposes, for evolving efficient systems of shorthand and other codes etc. etc. The interested reader may refer to the discussion in Dewey (1923) where valuable data on English are presented, to Herdan (1956, Section 5.4) where the relative frequencies of phonemes are used for studying the 'distances' between languages, and to Herdan (1956, Section 7.3) where an application to crytography is demonstrated. Shannon (1951) used such data for analysing the redundancy of written English from the information-thoretic point of view, Zipf (1949, pp. 97-109) used such estimates for expounding his views on general principles underlying human languages.

2. Of particular interest in the present context are the data on Sanskrit phonemes compiled by Whitney (1923, pp.10-26) and those on Bengali phonemes compiled by Chatterjee (1926, pp.270-4). Chatterjee makes very important observations on the shifts in the distribution from Sanskrit to Bengali.

3. Tables 1(a) and (b) present some statistics on relative frequencies of letters in modern Bengali prose. Table 1(a) shows the relative frequencies of letters in 'Visavriksha' and 'Sheser Kavita', and Table 1(b) gives some details for the letter 'a' (অ). The material used is the same as that employed in Appendix 1 for studying word-length in Bengali prose in terms of letters.

4. The silent a features of Tables 1(a) and 1(b) are given in the following paragraphs. It may be noted that the relative frequencies of Bengali phonemes estimated by Chatterjee (1926, pp. 270-4) [vide Chap. 1, section 1.5] lead to very similar conclusions. This is expected since the correspondence between phonemes and letters is fairly close in Bengali.

5. The most frequent letters are all vowels - অ (a) occurs with a frequency of nearly 15% (including over 2% silent অ's ; আ (ā) has a relative frequency between 10-11%, and ে (e) between 9-10%. The other letters may be distributed as follows on the basis of the average relative frequency in Visavriksha and Sheser Kavita :

percentage frequency	letters
less than 0.1	ৈ (ai) ৌ (au) ঞ (ñ) ঢ (dh) ঃ (ḥ) ষ (ṭ)
0.1 - 0.5	ূ (ū) ঋ (ri) ঘ (gh) ঙ (ṅ) ঝ (jh) ঠ (ṭh) ঢ (ḍ) ফ (ph) ণ (ṇ) ং (ṁ)
0.5 - 1	ী (ī) খ (kh) চ (ch) ছ (chh) জ (j) ট (ṭ) ন (n) থ (th) ধ (dh) ভ (bh) শ (sh) ষ (ṣ) হ (h) ড় (ḍ)
1 - 2	ো (o) গ (g) য (y)
2 - 3	ু (u) দ (d) প (p) ম (m) স (s) ় (y)
3 - 4	ত (t) ব (b, v) ল (l)
4 - 5	ক (k) ন (n)
5 - 6	- -
6 - 7	র (r) ি (i)

Table 1(a) Relative frequencies of different letters of the Bengali alphabet based on the probability samples of words* from "Visavriksha" and "Sheser Kavita".

sr. no.	letter	percentage frequency									
		Visavriksha					Sheser Kavita				
		ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	অ (a)	15.89	16.28	16.43	15.16	15.92	12.05	12.68	13.94	12.75	12.84
2	আ (ā)	9.94	8.77	10.88	9.40	9.74	11.70	11.00	9.30	10.66	10.69
3	ই (i)	7.25	7.74	7.70	7.00	7.42	6.18	4.98	5.73	6.15	5.75
4	ঈ (ī)	1.40	1.37	0.91	1.57	1.32	0.53	0.62	0.32	1.09	0.64
5	উ (u)	2.92	2.05	2.15	3.66	2.71	1.70	1.97	3.14	3.51	2.58
6	ঊ (ū)	0.70	0.68	0.45	0.73	0.64	0.32	0.10	0.22	0.11	0.19
7	ঋ (ri)	0.82	0.34	0.68	0.42	0.56	-	0.42	0.43	-	0.21
8	এ (e)	5.97	7.97	7.13	7.63	7.19	11.17	11.73	11.02	10.76	11.20
9	ঐ (ai)	0.23	0.11	0.11	0.10	0.14	0.11	-	0.11	-	0.05
10	ও (o)	0.82	1.48	0.57	0.52	0.84	1.60	2.49	1.95	2.20	2.06
11	ঔ (au)	-	0.11	0.11	0.21	0.11	0.11	-	-	0.22	0.08
	vowels (1-11)	45.94	46.90	47.12	46.40	46.59	45.47	45.99	46.16	47.45	46.29
12	ক (k)	4.91	3.99	4.98	4.81	4.67	4.90	5.30	5.18	4.29	4.93
13	খ (kh)	0.70	1.48	1.25	0.84	1.06	0.75	0.73	1.08	0.99	0.88
14	গ (g)	1.52	1.25	1.25	1.46	1.37	0.85	0.93	1.73	0.99	1.12
15	ঘ (gh)	-	-	0.11	-	0.03	0.11	0.31	0.11	0.33	0.21
16	ঙ (ṅ)	0.23	0.23	-	0.21	0.17	-	0.31	0.32	0.22	0.21
	sub-total (12-16)	7.36	6.95	7.59	7.32	7.30	6.61	7.58	8.42	6.82	7.35
17	চ (ch)	0.35	0.57	0.34	0.63	0.48	0.96	0.31	0.54	1.21	0.75
18	ছ (chh)	0.47	0.46	0.79	0.52	0.56	1.28	0.73	1.51	1.43	1.23
19	জ (j)	0.70	0.23	0.34	0.31	0.39	1.28	1.25	0.76	1.65	1.23
20	ঝ (jh)	-	-	0.34	0.21	0.14	-	0.10	-	0.33	0.11
21	ঞ (ñ)	-	-	-	0.21	0.06	0.11	0.31	-	0.11	0.13
	sub-total (17-21)	1.52	1.26	1.81	1.88	1.63	3.63	2.70	2.81	4.73	3.45

(contd.)

Table 1(a) : (Contd.)

sr. no.	letter	percentage frequency									
		Visavriksha					Sheser Kavita				
		ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
22	ন্ (t)	0.47	0.11	0.23	0.31	0.28	0.85	1.35	1.62	0.99	1.20
23	ন্থ (th)	-	-	0.11	0.10	0.06	0.32	0.21	0.22	0.55	0.32
24	ন্দ (d)	-	-	-	0.21	0.06	0.11	0.21	0.22	0.22	0.19
25	ন্ধ (dh)	-	-	0.23	-	0.06	-	-	-	-	-
26	ন্ (n)	0.59	0.46	0.68	0.94	0.67	0.43	0.21	0.22	0.55	0.35
sub-total (22-26)		1.06	0.57	1.25	1.56	1.13	1.71	1.98	2.28	2.31	2.06
27	ত (t)	4.33	5.13	4.19	3.55	4.28	3.83	3.43	4.11	3.41	3.69
28	ৎ (t)	-	-	-	0.21	0.06	-	-	0.22	-	0.05
29	থ (th)	0.47	0.46	1.02	0.63	0.64	0.75	1.04	0.76	0.22	0.70
30	দ (d)	2.34	3.64	3.17	2.93	3.02	1.81	2.08	2.49	1.87	2.06
31	ধ (dh)	0.35	0.80	0.79	1.04	0.76	1.06	0.52	0.54	0.33	0.61
32	ন (n)	5.38	6.61	4.54	6.27	5.74	4.15	3.32	4.00	4.94	4.09
sub-total (27-32)		12.87	16.64	13.81	14.63	14.50	11.60	10.39	12.12	10.77	11.20
33	প (p)	2.11	2.62	2.49	1.25	2.10	2.56	2.28	2.05	1.87	2.19
34	ফ (ph)	0.23	-	0.11	-	0.08	0.11	0.21	0.11	0.33	0.19
35	ব (b,v)	2.81	2.62	2.83	3.34	2.91	4.58	4.26	3.78	3.19	3.96
36	ভ (bh)	0.35	0.46	0.23	0.31	0.34	0.85	0.73	0.32	0.77	0.67
37	ম (m)	2.92	2.73	2.83	2.30	2.69	2.66	3.43	3.03	3.63	3.18
sub-total (33-37)		8.42	8.43	8.49	7.20	8.12	10.76	10.91	9.29	9.79	10.19
38	য (y)	1.29	1.25	1.47	1.78	1.46	1.81	1.56	1.30	1.21	1.47
39	র (r)	7.25	7.18	6.57	6.06	6.75	7.87	6.96	6.48	6.59	6.99
40	ল (l)	3.51	2.62	2.38	2.51	2.74	2.98	3.63	3.57	3.30	3.37
sub-total (38-40)		12.05	11.05	10.42	10.35	10.95	12.66	12.15	11.35	11.10	11.83

(contd.)

Table 1(a) (contd.)

sr. no.	letter	percentage frequency									
		Visavriksha					Sheser Kavita				
		ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
41	শ (sh)	1.29	0.34	0.23	0.84	0.67	0.75	1.14	0.43	1.09	0.86
42	স (s)	1.05	0.57	0.68	0.42	0.67	0.32	0.52	0.54	0.44	0.45
43	ষ (s)	2.57	2.28	2.04	3.66	2.66	2.45	2.80	2.38	2.42	2.52
44	হ (h)	1.99	2.62	2.26	2.82	2.43	0.53	0.93	0.76	0.55	0.70
sub-total (41-44)		6.90	5.81	5.21	7.74	6.43	4.05	5.39	4.11	4.50	4.53
45	য় (y)	2.69	1.48	2.04	2.51	2.18	1.92	1.87	2.27	2.09	2.03
46	ড় (ḍ)	0.59	0.46	1.13	-	0.53	0.53	0.73	0.22	0.44	0.48
47	ঢ় (ḍh)	-	-	-	-	-	-	-	-	-	-
48	ং (ṅ)	0.23	0.23	0.11	-	0.14	0.21	0.21	0.52	-	0.19
49	ঃ (h)	0.12	0.11	0.11	0.10	0.11	-	-	0.11	-	0.03
50	ঁ (n)	0.23	0.11	0.91	0.31	0.39	0.85	0.10	0.54	-	0.37
total (1-50)		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
silent অ (a)		2.33	2.51	2.84	2.09	2.43	1.50	1.87	2.38	2.86	2.14
no. of sample words		151	146	147	167	611	188	178	185	184	735
no. of sample letters		854	877	882	957	3570	939	963	925	910	3737

* that is, words falling on 100 randomly selected lines from each work.

Table 1(b) Relative frequencies of the letter অ (a) in different positions within words based on the probability samples of words* from "Visavriksha" and "Sheser Kavita".

position etc.	percentage frequency									
	Visavriksha					Sheser Kavita				
	ss 1	ss 2	ss 3	ss 4	comb.	ss 1	ss 2	ss 3	ss 4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
initial (pronounced)	0.47	0.11	0.45	0.52	0.39	0.32	0.73	0.54	0.66	0.56
medial (pronounced)	10.42	10.24	10.54	10.04	10.30	8.63	8.52	9.62	7.14	8.48
medial (silent)	0.58	0.57	0.57	0.31	0.50	0.75	0.62	0.54	0.55	0.61
final (pronounced)	2.69	3.42	2.60	2.51	2.80	1.60	1.56	1.40	2.09	1.66
final (silent)	1.75	1.94	2.27	1.78	1.93	0.75	1.25	1.84	2.31	1.53

* that is, words falling in 100 randomly selected lines from each work.

6. The vowels form about 46% of all letter-occurrences. Among the 25 stops or occlusives, the gutturals, the dentals and the labials are more frequent and the palatals and cerebrals relatively rare.

7. Frequency is larger for the short इ (i) than for the long ई (ī), and similarly for the short उ (u) than for the long ऊ (ū); the two diphthongs ऐ (ai) and औ (au) are much rarer than the vowel ओ (o). The long आ (ā) is, however, not far short of अ (a).

8. The aspirate letters are rare in comparison with the corresponding inaspirate letters; छ (chh) is a notable exception. Generalisation is difficult for the 'ghosa' (voiced) vs 'aghosa' (voiceless) comparisons, for many voiced letters (eg., j, dh, b, bh) are more frequent than their voiceless counterparts.

9. Table 1(b) shows that the silent a's (अ) are mostly at the final position, and among final a's a high proportion are silent.

10. If one compares the relative frequencies for Visavriksha and Sheser Kavita - the former is written in a more Sanskritised language than the latter where 'tadbhava' (Prakrit) forms are more frequent - one finds that the relative frequencies are particularly different for vowels. Sheser Kavita uses अ (a), इ (i), ई (ī), ऋ (ri) and ऊ (ū) less frequently, and ओ (o), and ए (e) more frequently than in Visavriksha. Among consonants, Sheser Kavita uses घ (gh), छ (chh), ज (j), ट (ṭ), ठ (ṭh), ब (b) more frequently and

ण (ṇ), ह (h) and ः (ḥ) less frequently. Some of the differences are conspicuous. Broadly speaking, the differences can be explained by considerations similar to those put forward by Chatterjee (*ibid*) when comparing relative frequencies of Sanskrit and Bengali phonemes.

Appendix 5 : The rank-frequency relation for words in Bengali prose

1. The rank-frequency relation for words is among the most famous findings of quantitative linguistics. Briefly, the finding is this : If a word-count is carried out on a sufficiently long text and the frequencies of different words occurring in the text determined, the frequencies of the different words are found to follow the harmonic progression, approximately. If f_r is the frequency of the rth commonest word, then $f_r = \frac{c}{r}$ approximately ($r = 1,2,3,....$). If f_r is plotted against r on double-logarithmic scale, the relationship is approximately linear with a slope of minus 1. This relation is approximately equivalent to a Pareto distribution for the variate word-frequency (f_r).

2. First observed by Estoup and Condom, this relation was discussed in detail by Zipf (1949), who saw in it an expression of the universal principle of least effort. The relation has been observed for a wide variety of languages, for texts written by many different writers writing on diverse types of subjects. Various theories or probability models have therefore been put forward to throw light on its genesis. Among them may be mentioned the work of Mandelbrot (1953. 1954), Miller (1958), Miller and Newman (1958), Miller, Newman and Friedman (1958), Simon (1955), Good (1957), and Herdan (1957, 1960, 1961)[1/].

3. Tables 1(a) and (b) present the results of two word-counts on modern Bengali prose. Table 1(a) is based on a complete count on

---

1/ Vide Chapter 1, Section 1.3, for a more detailed discussion.

'Gora', the well-known (about 600-page) novel by Tagore, written in what may be called a model of Bengali prose, excepting that the non-conversational matter uses the chaste ("Sadhu") forms of verbs and pronouns[1]. Table 1(b) is based on a count covering the first (nearly 30-page) story in the collection of stories entitled "Chacha Kahini" by Muztaba Ali. This work is written in the colloquial style and the vocabulary contains a fair proportion of words of European and Mohammedan origin. In both the counts, each inflected form of a given root word was kept separate. This was in accordance with standard linguistic practice, as recommended by Zipf (1949, p. 23) and exemplified in the excellent "Word-index to Ulysses" by Hanley (1937)[2].

4. Figure 1 shows the rank-frequency relation observed in the two word-counts. Following Zipf (1949), the frequency f_r of the word ranked r in descending order of frequency is plotted against the rank r, and both variables are used in the logarithmic form. Where some words have the same frequency, we get short horizontal steps in the diagram. [The modified procedure of Miller, Newman and Friedman (1958) was not followed : This modification consists in plotting the common frequency against the average rank, so that the horizontal steps do not appear.] A straight line with slope-1 has been placed on the figure to help in visual judgment.

1/ The author is indebted to Shri Jibanmoy Roy of the Linguistic Research Unit of the ISI for making the data available to him. Roy's count showed many group verbs (e.g., "chaliya gela") as single words; the present author did the work needed for splitting these groups and adjusting the word-frequencies accordingly.

2/ Yule (1944, pp.32-33) criticises this approach and his count is almost based on the root-word approach. The difference is not very serious for English nouns considered by Yule.

Table 1 Frequency distribution of words by number of occurrences in "Gora" and "Chacha Kahini" (first story)

(a) Gora

no. of occurrences (f)	no. of words	no. of occurrences (f)	no. of words	no. of occurrences (f)	no. of words	no. of occurrences (f)	no. of words
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	7818	31	17	61	3	94	3
2	2219	32	24	62	8	95	2
3	1197	33	15	63	3	97	2
4	740	34	13	64	4	99	2
5	488	35	20	65	7	100	2
6	320	36	12	66	2	101	1
7	268	37	11	67	1	102	3
8	225	38	7	68	6	104	1
9	161	39	10	69	4	105	4
10	134	40	11	70	7	107	2
11	122	41	7	71	1	108	3
12	98	42	10	72	3	109	3
13	81	43	14	73	3	112	2
14	89	44	5	74	4	113	3
15	76	45	5	75	4	114	2
16	49	46	5	76	3	115	1
17	46	47	11	77	4	116	2
18	66	48	9	78	2	117	1
19	37	49	10	79	6	118	1
20	41	50	15	80	5	120	2
21	45	51	8	81	4	121	2
22	33	52	8	82	4	122	1
23	23	53	12	83	3	124	1
24	24	54	4	85	2	125	2
25	22	55	10	86	6	127	1
26	19	56	5	88	4	128	3
27	19	57	11	90	2	130	1
28	20	58	4	91	2	131	1
29	16	59	2	92	2	132	1
30	10	60	3	93	1	135	1

Table 1 : (Contd.)

(a) Gora

no. of occurrences (f) (1)	no. of words (2)	no. of occurrences (f) (1)	no. of words (2)	no. of occurrences (f) (1)	no. of words (2)	no. of occurrences (f) (1)	no. of words (2)
136	1	185	2	281	1	505	1
137	2	186	1	304	1	515	2
139	3	188	1	337	1	520	1
140	2	191	1	338	1	531	1
141	1	196	1	339	1	540	1
142	1	197	1	344	1	541	1
147	1	199	1	347	1	546	1
148	1	201	1	349	1	560	1
150	2	202	1	352	1	562	1
152	1	204	2	355	1	578	1
154	3	212	1	356	1	626	1
155	1	214	1	360	1	654	2
156	1	216	1	363	2	697	1
157	1	220	1	370	1	716	1
160	1	225	1	376	1	775	1
163	1	226	1	377	1	830	1
164	1	227	2	391	1	888	1
165	1	228	1	394	1	904	1
168	2	235	1	402	1	961	1
169	1	237	1	411	1	974	1
171	1	246	1	435	1	978	1
172	1	250	1	448	1	995	1
173	1	255	1	468	2	1085	1
174	1	257	2	472	2	1119	1
175	1	258	1	483	1	1164	1
177	1	262	1	485	1	1404	1
178	1	268	1	487	1	1489	1
180	1	272	1	499	1	1516	1
181	1	276	1	500	1	2945	1
183	1	278	1			total	1505

total no. of word-occurrences : 126359

Table 1 : (Contd.)

(b) Chacha Kahini (first story)

no. of occurrences (f) (1)	no. of words (2)	no. of occurrences (f) (1)	no. of words (2)
1	1233	20	1
2	262.	24	2
3	118	26	2
4	55	27	1
5	35	28	1
6	23	29	3
7	21	30	1
8	18	31	1
9	16	32	2
10	4	33	1
11	5	34	3
12	5	48	1
13	4	50	1
14	2	51	1
15	4	61	1
16	2	63	1
17	4	96	1
18	4	total	1840
19	1		

total no. of word-occurrences : 4456

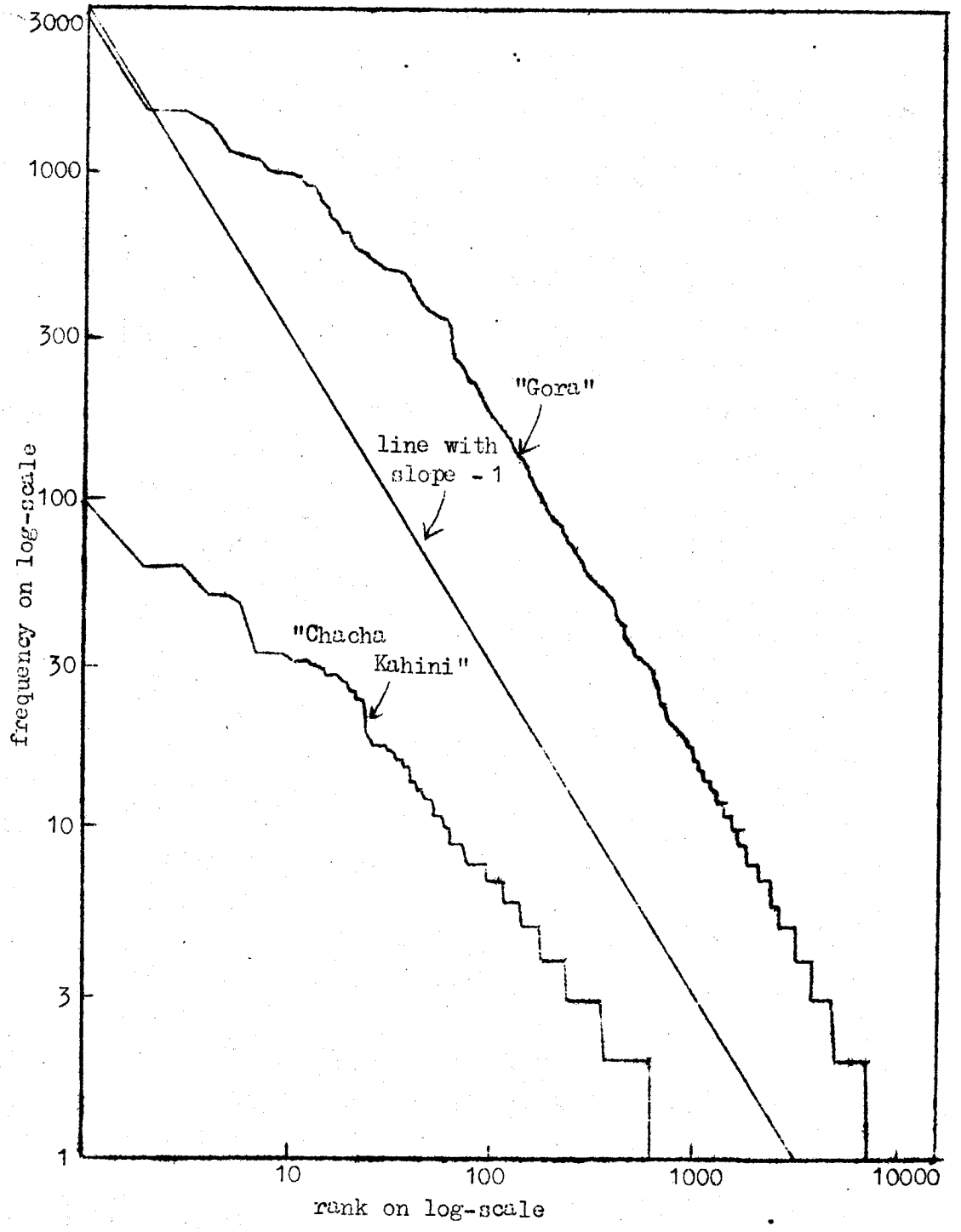


Fig.1: Rank-frequency relation for all words based on a complete count of (i) "Gora" - 15105 distinct words, 126359 occurrences and (ii) the first story of "Chacha Kahini" - 1840 distinct words, 4456 occurrences.

5. The curve for 'Chacha Kahini' is sensibly straight with slope nearly -0.8 above log (rank) nearly 1.0, but below this the curve shows some irregularity and top-downward concavity. The curve for 'Gora' looks nearly straight above log (rank) about 1.5 with a slope slightly steeper than -1, and below this the curve displays a very clear downward concavity. Apparently the rank-frequency relation is not very clear in Bengali prose.

6. A closer study of Zipf's (1949) work, however, reveals that similar discrepancies have been observed in other cases. We may mention three points mentioned by Zipf as particularly relevant. First of all, it is recognised that the slope may increase somewhat as the sample size increases. This is intuitively evident. For as the word count is enlarged, the frequencies of the commonest words tend to increase proportionately, but the total number of word-types would tend to increase, though less than proportionately, so that the frequencies of the rare words would increase less than proportionately. Also the curve may be erratic in small samples. This may explain why the curve for 'Chacha Kahini', admittedly based on a small sample, looks rather erratic and also has a slope appreciably below 1. [cf. curve B of Fig. 3.15 on p. 125 of Zipf (1949); based on 1726 words of old High German, this curve shows a slope near -0.85.]

7. Secondly, the top downward concavity has been observed very frequently near the first dozen or so most frequent words; this, it is believed, is a general characteristic of informal writing or

colloquial speech, (e.g. of letters) where articles like 'the' are frequently omitted.

8. Finally, for holophrasic[1] languages like Hebrew, Plains Greek or Nootka — the two latter-mentioned are American Indian languages — the rank frequency curve has been found to have slopes appreciably below 1. This, it is agreed, is due to the fact that the more frequent words will tend to be used in a larger number of permutations. For Nootka, the r-f relation after splitting holophrases into constituent parts appeared to be close to standard form, while the r-f relation before splitting holophrases is considerably flatter. This point may be important for Bengali which uses a good proportion of compounds.

9. So far as the count on 'Gora' is concerned, there is yet another reason why the r-f relation may deviate from the standard one. 'Gora' uses chaste forms of verbs, pronouns etc. in the non-conversational matter and the colloquial forms of the same words in speeches by different characters. This splits the total frequency of these common grammar forms into two, so that either the chaste form or the colloquial form may not be very frequent. This might explain why the top-downward concavity extends upto several dozen words.

1/ Holophrases are permutations of two or more words, like 'brother-in-law' which behave like single words so far as inflected forms are concerned.

Appendix 6 : An experimental comparison between speed of writing in English and Bengali

1.1. Introduction : To a statistically-minded person, one of the major problems of linguistics would be to examine, at least roughly, the relative efficiencies or economies of different languages considered as different systems of codes for conveying information. For comparing two languages in their spoken forms, one should, in principle, consider pairs of "equivalent" passages; the two passages in each pair should be in the two languages, but should express exactly the same message or idea. One should then compare the lengths of the two passages of each pair, length being measured by the number of phonemes (not syllables). Strictly speaking, even the number of phonemes cannot serve as a perfect basis for comparisons, for different phonemes may require different amounts of time and energy for articulation [vide Miller, 1951, pp. 26-41 ; Chatterjee, 1921 , xxxxxxxx Zipf, 1949, pp.97-109]; these aspects should also be taken into account.

1.2. Written forms of languages are, from the linguistic point of view, far less important than spoken forms [Taraporewala, 1951, pp. 5-6]. Still, there would be much interest in comparing the efficiencies of two languages in their written forms. Here again, one should not depend entirely on the number of letters (say) for purposes of comparison; one should really consider the time and energy spent in writing equivalent passages in the different languages.

1.3. In this Appendix is reported an experiment carried out for answering the following question : If the same message is conveyed

by two passages, one in English and the other in Bengali, and if both passages are written by the same subject exerting himself to the same extent and under the same conditions, what would be the comparative time requirements? The experiment was on a small scale, and all five subjects had Bengali as their mother tongue; what is more, there were certain insuperable theoretical difficulties which made really scientific experimentation impossible. Nevertheless, it seems possible to conclude that the time requirements for written expression are roughly of the same order so far as English and Bengali are concerned.[1/]

1.4. The reader is referred to Chapter 1, Section 1.4, where previous studies on comparative efficiencies of languages are reviewed. It is proposed to compare, in a later communication, the efficiencies of English and Bengali by using statistical data on lengths of passages and information-theoretic and other approaches.

1.5 Very little seems to have been done on these lines, especially for written forms of languages. It is no doubt difficult to compare the spoken forms [vide para 1.1 Supra], but as regards written forms, rough comparisons can be easily made as reported in the present Appendix. Similar work can be done on comparative time requirements for typing, printing etc. in different languages. And if some alternative system of writing is being proposed for any language, e.g., the romanised system for Bengali [vide S. K. Chatterjee, 1935], one can compare the original and the proposed scripts in this manner.

1/ This is rather contrary to popular impressions, according to which the Bengali script is appreciably less efficient than the English one. The contradiction will be explained in Section 5.

2.1 The experimental material : The first step was to select pairs of equivalent passages. An ideal pair of equivalent passages would be such that each passage of the pair is the unique and perfect (i.e. completely faithful) translation of the other. (Without the properties of uniqueness and faithfulness, the question sought to be answered through the experiment becomes somewhat imprecise.) Passages satisfying these criteria were not easily available; for translations are too often of the "free" type. Anyway, the following 27 passage-pairs were finally selected for the experiment :

(a) Six passages were chosen from "Shesher Kavita", a Bengali novel by Tagore. Three of these passages are conversational. The corresponding English passages were taken from the English version, "Farewell My Friend", by K. R. Kripalani. The language of the original is highly idiomatic, but the translation seemed to be quite faithful.

(b) Ten English passages were selected from a larger number set in different years as passages for translation into Bengali in the Bengali Second Paper of the School Final Examination, Board of Secondary Education, West Bengal. The 'model' Bengali translations were taken from a text-book on Bengali composition ("Nava Praveshika Rachana O Anuvad" by Ashok Shastri and Shashankashekhar Bagchi, Modern Book Agency, Calcutta). Some of the English passages were really translations of Bengali originals. Generally, these passages covered elementary topics (stories etc.) of a more or less factual nature.

(c) Six passages were selected from Jawaharlal Nehru's "Autobiography" and the corresponding Bengali passages from the translation by Satyendranath Majumdar.

(d) Five passages on different political topics were extracted from news published in the English dailies of Calcutta. These were emanating from Reuter and other agencies. The Bengali versions appeared in the Bengali dailies of Calcutta, but the translation had to be brushed up a little in order to make them more faithful to the originals.

2.2. The lengths of these passages are shown in cols.(3) and (4) of Table 1. Generally, all the passages were chosen to be short, so that effects of fatigue would not appear. Passages of set (b) were taken whole; even then they are generally shorter than the others.

2.3. Five persons DPB, SC, PB, RM and NB (the present author), participated as subjects in the writing experiment. Each subject wrote out each of the 27 pairs of passages. The pairs were taken up in the same order. ~~In every case, the time required was recorded in the same order~~ In every case, the time required was recorded in seconds by an observer. For any subject and for any pair of passages, the two passages were written on the same day consecutively, with only a few minutes' rest in between, under the same conditions(pen, posture etc.); and the subject tried to put in equal efforts for writing both the passages. If for any pair, the English passage was written first and the Bengali passage next (by all subjects), the

order was reversed (for all subjects) for the next pair of passages. This was expected to compensate for fatigue or practice, if any. But the idea was not pursued far; thus, the subjects sometimes wrote more than one passage-pair on the same day, which should perhaps have been avoided.

2.4. Generally, the subjects were advised to write fast, as the equalisation of effort seemed to be easier when the subjects were writing fast. But it was expressly laid down that each letter of the writing must be legible, that is, there should not be scrawls or waves in the writing where the reader has to exercise some judgement. Actually, for psychological reasons, DPB and NB chose to write as fast as possible, even to the detriment of quality, while the others wrote fairly good hands; RM and PNB wrote slower than DPB or NB, and SC wrote at a still slower pace. But even SC was writing at fair speed.

3.1. Results : Table 1 shows the data thrown up by the experiment. For each pair of equivalent passages, this table shows : (1) the number of words in the English and in the Bengali passages, and (ii) the time required by each subject for writing each passage of the pair. Appropriate sub-totals are inserted to facilitate examination.

3.2. Table 2 is based on Table 1. It gives, for convenience of study, the ratios between the two time requirements, Bengali : English, expressed as percentages. These ratios are given by subjects and by passage-pairs. The sub-total figures in Table 1 were used for obtaining the ratios in the sub-total rows and columns of Table 2; the latter are not simple means of the individual ratios.

Table 1: Comparative time requirements for copying equivalent passages in English and in Bengali, separately by passage-pairs and by subjects.

source/type of passage-pair	sr. no.	length (no. of words)		time required in seconds by subjects (initials shown) for each passage of the pair.											
				DPB		SC		PNB		RM		NB		all subjects	
		English passage	Bengali passage	English	Bengali	English	Bengali	English	Bengali	English	Bengali	English	Bengali	English	Bengali
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
"Sheser Kavita", a	1	262	213	466	368	828	634	712	700	627	474	520	450	3153	2626
Bengali novel by Tagore,	2	240	162	498	346	692	554	643	530	476	395	438	412	2747	2237
and its English trans-	3	213	159	386	293	486	422	453	334	370	314	350	309	2045	1672
lation	4	299	201	523	409	701	549	567	484	539	463	471	407	2801	2312
	5	253	206	388	341	624	535	421	403	475	413	377	373	2285	2065
	6	383	266	665	496	828	676	603	572	663	582	540	512	3299	2838
sub-total	1-6	1650	1207	2926	2253	4159	3370	3399	3023	3150	2641	2696	2463	16330	13750
English passages	7	110	97	158	153	210	200	160	173	164	186	149	168	841	880
set in a public	8	82	64	119	104	150	142	108	112	122	121	107	107	606	586
examination for	9	122	90	191	167	260	232	165	193	191	200	164	169	971	961
translation into	10	101	83	153	142	195	192	161	166	138	158	138	145	785	803
Bengali, and the	11	82	74	129	120	178	197	138	168	151	170	118	146	714	801
model answers	12	104	90	155	152	203	227	166	181	147	181	134	160	805	901
	13	133	119	173	220	311	313	224	253	236	280	198	214	1142	1280
	14	124	96	200	200	271	269	204	218	214	295	178	188	1067	1170
	15	122	90	221	185	315	262	215	202	233	228	184	186	1168	1063
	16	146	121	228	214	324	295	225	228	255	272	217	227	1249	1236
sub-total	7-16	1126	924	1727	1657	2417	2329	1766	1894	1851	2091	1587	1710	9348	9681

(Contd.)

Table 1 (Contd.)

source/type of passage-pair	sr. no.	length (no. of words)		time required in seconds by subjects (initials shown) for each passage of the pair											
				DIB		SC		INB		RI		NB		all subjects	
		English passage	Bengali passage	English	Bengali	English	Bengali	English	Bengali	English	Bengali	English	Bengali	English	Bengali
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Nehru's "Autobiography" and its Bengali translation	17	218	160	394	339	539	478	404	402	413	423	351	360	2101	2002
	18	265	194	515	403	702	606	490	446	555	512	448	425	2710	2392
	19	246	197	471	375	592	571	463	439	472	475	411	416	2409	2276
	20	338	275	611	531	822	820	635	664	638	619	520	561	3226	3195
	21	223	212	386	405	494	588	407	509	424	497	355	420	2066	2419
	22	189	203	396	384	541	561	456	442	434	483	375	423	2202	2293
sub-total	17-22	1479	1241	2773	2437	3690	3624	2855	2902	2936	3009	2460	2605	14714	14577
(English) newspassages from Reuter etc. and the Bengali translation	23	108	91	204	186	277	258	216	236	207	223	193	206	1097	1109
	24	170	138	296	290	369	380	319	357	302	323	281	291	1567	1641
	25	176	134	288	291	370	410	327	359	313	352	281	295	1579	1707
	26	187	172	308	325	408	442	328	380	324	383	292	333	1660	1863
	27	161	128	268	252	331	342	271	306	285	291	252	260	1407	1451
sub-total	23-27	802	663	1364	1344	1755	1832	1461	1638	1431	1572	1299	1385	7310	7771

Table 2: Time requirement for copying a Bengali passage expressed as percentage of the time requirement for copying the equivalent English passage, separately by passage-pairs and by subjects.

source/type of passage-pair	sr. no.	relative time requirements, Bengali : English, as percentages, by subjects (initials shown)					
		DPB	SC	PNB	RM	NB	all subjects
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
"Sheser Kavita", a Bengali novel by Tagore, and its English translation	1	79.0	76.6	98.3	75.6	86.5	83.3
	2	69.5	80.1	82.4	83.0	94.1	81.4
	3	75.9	86.8	73.7	84.9	88.3	81.8
	4	78.2	78.3	85.4	85.9	86.4	82.5
	5	87.9	85.7	95.7	86.9	98.9	90.4
	6	74.6	81.6	94.9	87.9	94.8	86.0
sub-total	1-6	77.0	81.0	88.9	83.8	91.4	84.2
English passages set in a public examination for translation into Bengali, and the model answers.	7	96.8	95.2	108.1	113.4	112.8	104.6
	8	87.4	94.7	103.7	99.2	100.0	96.7
	9	87.4	89.2	117.0	104.7	103.0	99.0
	10	92.8	98.5	103.1	114.5	105.1	102.3
	11	93.0	110.7	121.7	112.6	123.7	112.2
	12	98.1	111.8	109.0	123.1	119.4	111.9
	13	127.2	100.6	112.9	118.6	108.1	112.1
	14	100.0	99.3	106.9	137.9	105.6	109.7
	15	83.7	83.2	94.0	97.9	101.1	91.0
	16	93.9	91.0	101.3	106.7	104.6	99.0
sub-total	7-16	95.9	96.4	107.2	113.0	107.8	103.6
Nehru's "Autobiography" and its Bengali translation	17	86.0	88.7	99.5	102.4	102.6	95.3
	18	78.3	86.3	91.0	92.3	94.9	88.3
	19	79.6	96.5	94.8	100.6	101.2	94.5
	20	86.9	99.8	104.6	97.0	107.9	99.0
	21	104.9	119.0	125.1	117.2	118.3	117.1
	22	97.0	103.7	96.9	111.3	112.8	104.1
sub-total	17-22	87.9	98.2	101.6	102.5	105.9	99.1
English news-passages from Reuter etc. and the Bengali translations	23	91.2	93.1	109.3	107.7	106.7	101.1
	24	98.0	103.0	111.9	107.0	103.6	104.7
	25	101.0	110.8	109.8	112.5	105.0	108.1
	26	105.5	108.3	115.9	118.2	114.0	112.2
	27	94.0	103.3	112.9	102.1	103.2	103.1
sub-total	23-27	98.5	104.4	112.1	109.9	106.6	106.3

3.3. The scatter diagrams(Figs. 1 to 4)present the data of Table 1. Each diagram relates to one set of passages, and each dot in a diagram represents one particular passage-pair written by one particular subject. The initials of the subject and the serial number of the passage-pair [col. (2) of Table 1] are written against the point. The x-coordinate represents the time-requirement for copying the English passage and the y-axis that for the Bengali passage of the pair. Two straight lines are passed through each scatter diagram. One is the equiangular line y=x, for which the two co-ordinates are equal. The other straight line passing through these diagrams will be explained below.

3.4. The scatter diagrams seem to indicate that the experimental data are neither too erratic nor too regular. Indeed, the scatter diagrams have a very reasonable type of appearance.

3.5. Figures 5 and 6 are added to show that for passages of the size included in the present study, the effects of fatigue or practise are negligible. Fig. 5 shows the (average) time required (by the five subjects) for copying the 27 English passages against the lengths of the passages in terms of number of words; Fig. 6 is completely parallel, but relates to the Bengali passages. In either case, the regression seems to be a straight line passing through the origin; the constant of proportionality is about 32 words per minute for the English passages and 27 words per minute for the Bengali ones[1]. There seems to be some small variation between different sets of passages in regard to this constant of proportionality.

1/ NB wrote about 38 English wrods per minute and SC about 25 words; DPB wrote nearly 31 Bengali words per minute and SC about 22 words. Figs. 5 and 6 show the average speed of five subjects.

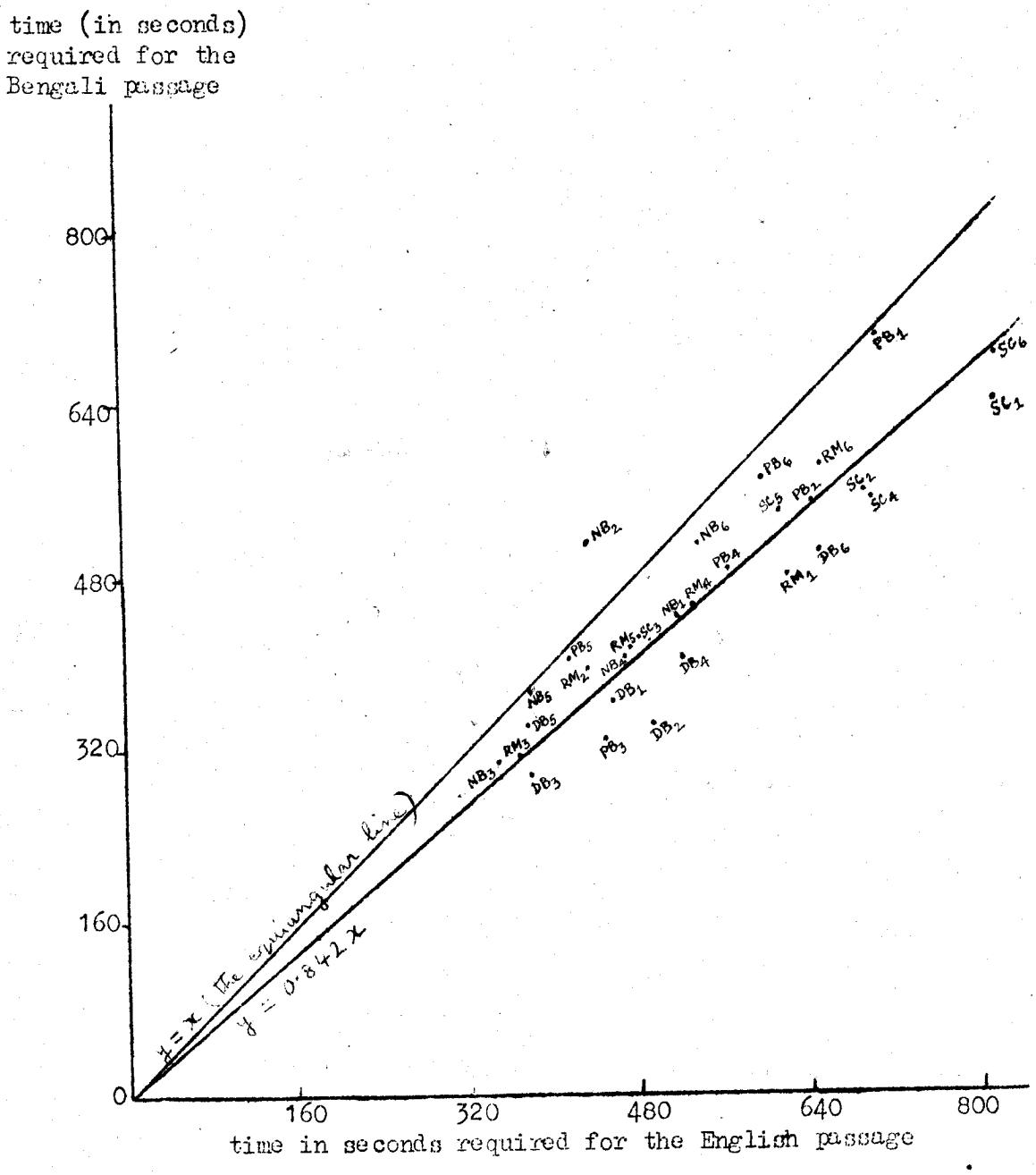


Fig. 1 : Comparative time requirements for copying selected passages from "Sheser Kavita" and corresponding passages from its English translation, separately by passage-pairs (serial nos. shown against each point) and by subjects (initials shown against each point).

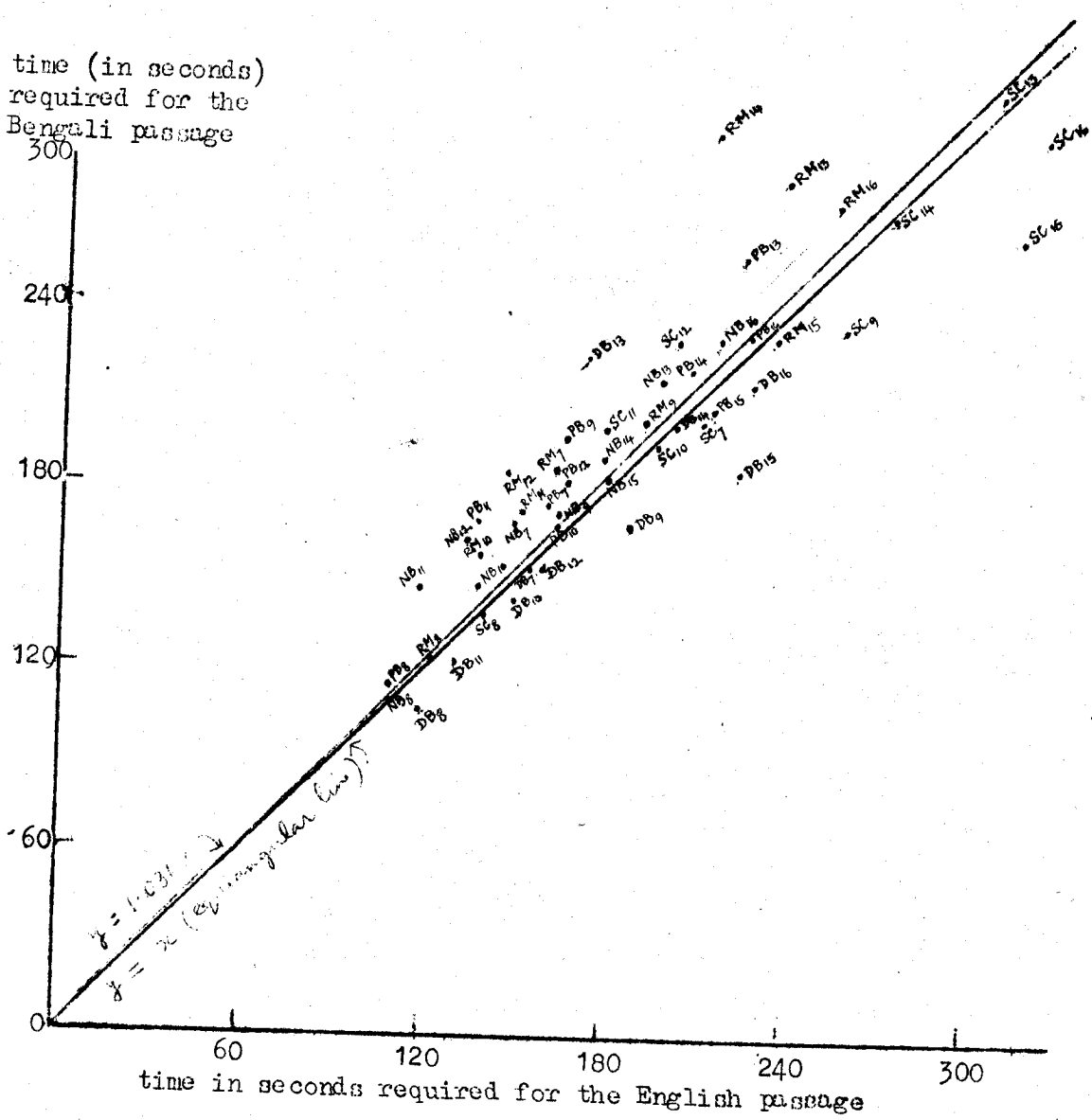


Fig. 2 : Comparative time requirements for copying selected English passages set in a public examination for translation into Bengali and the "model" answers, separately by passage-pairs (serial nos. shown against each point) and by subjects (initials shown against each point).

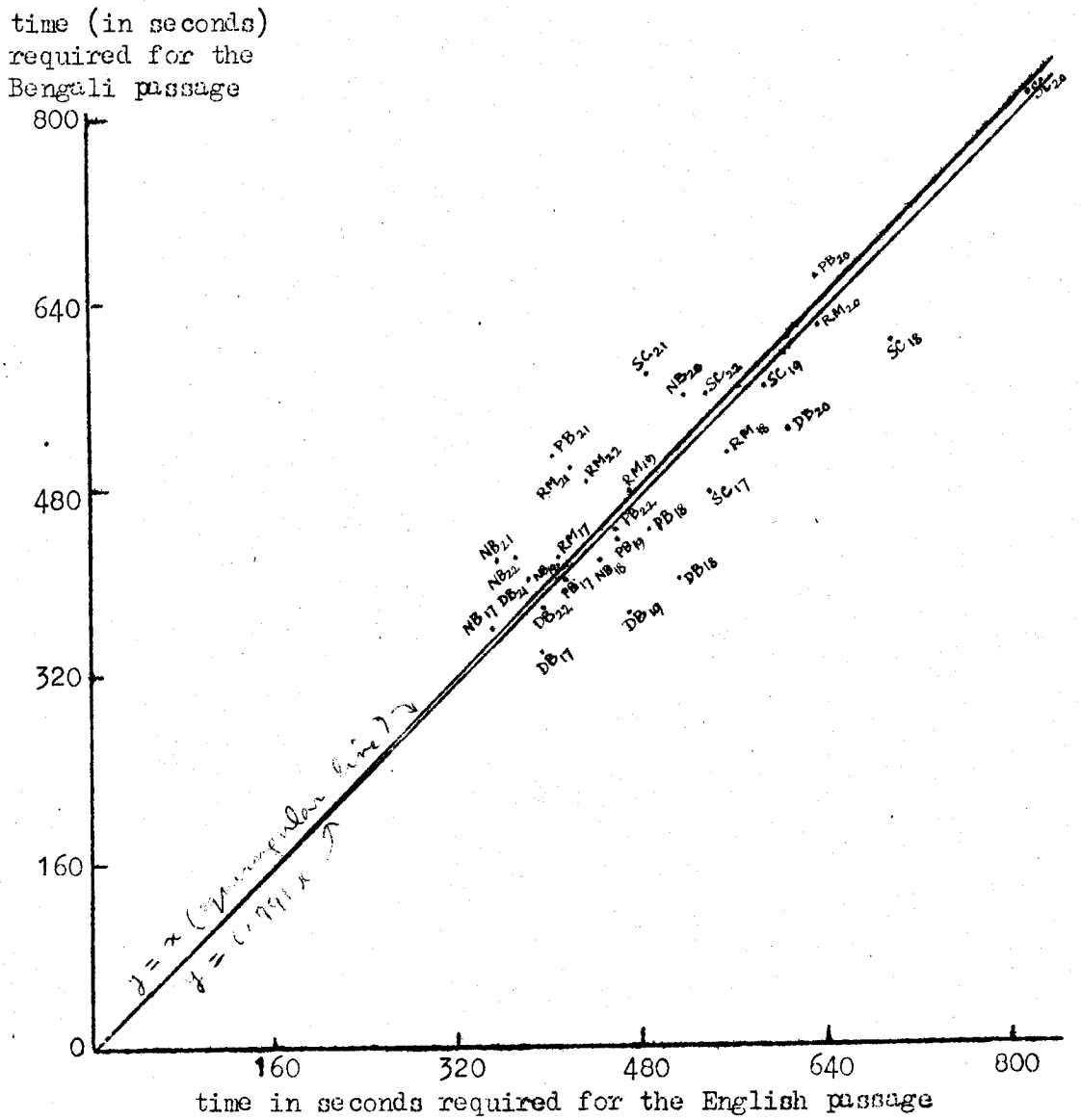


Fig. 3 : Comparative time requirements for copying selected passages from Jawaharlal Nehru's "Autobiography" and corresponding passages from its Bengali translation, separately by passage-pairs (serial nos. shown against each point) and by subjects (initials shown against each point).

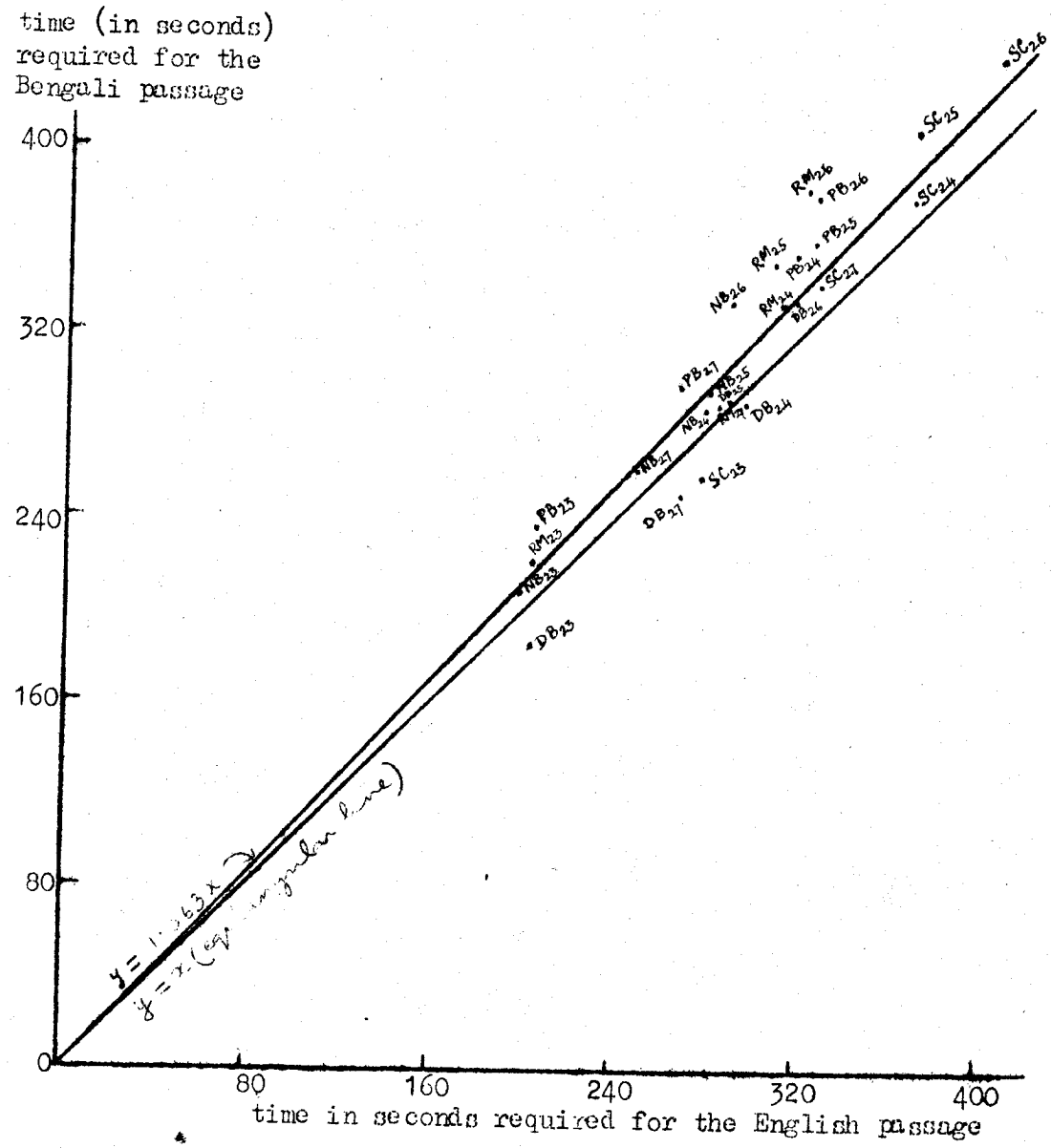


Fig. 4 : Comparative time requirements for copying selected newspassages emanating from Rauter etc., and their Bengali versions, separately by passage-pairs (serial nos. shown against each point) and by subjects (initials shown against each point).

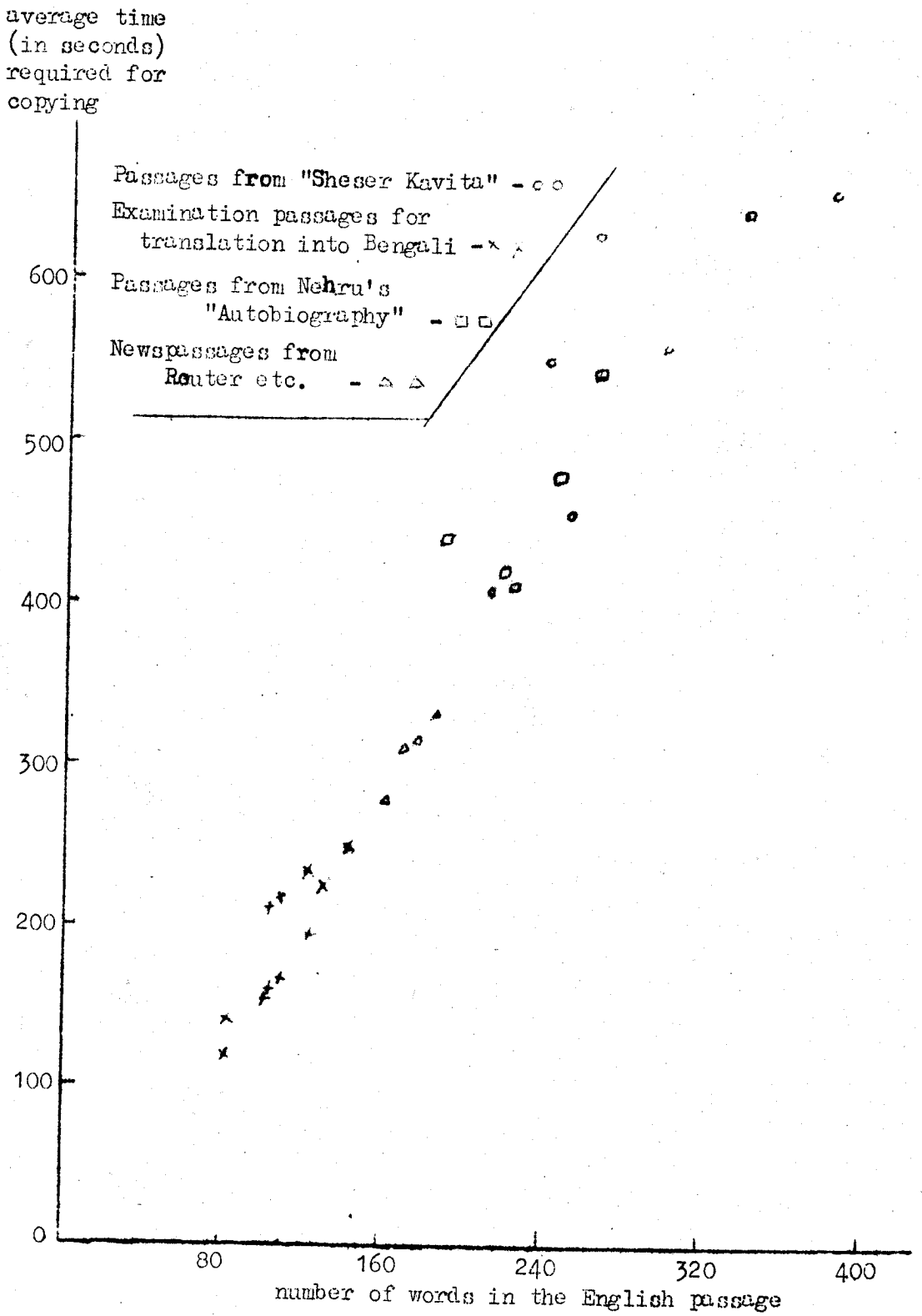


Fig. 5 : Average time required by five subjects for copying the 27 English passages against the lengths of passages in terms of words.

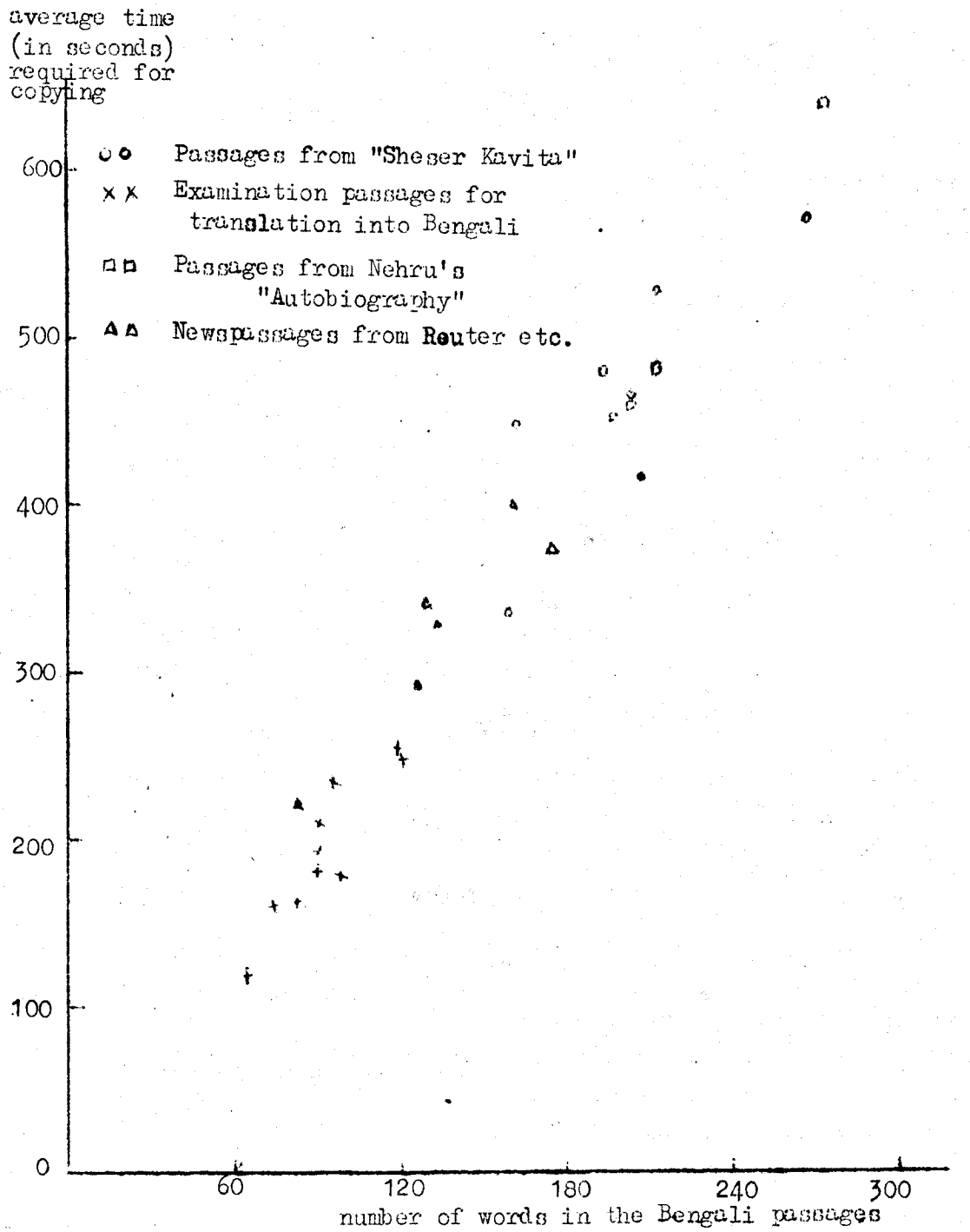


Fig. 6 : Average time required by the five subjects for copying the 27 Bengali passages against the lengths of the passages in terms of words.

3.6. Table 2 and the scatter diagrams (Figures 1 to 4) reveal many interesting points. Elaborate tests of significance seem to be hardly necessary.

3.7. First, there is some variation between subjects in respect of the relative time requirements. On the whole, DPB showed the lowest Bengali : English ratios, followed by SC, while the other three showed nearly equal ratios, on the average. The difference between DPB and NB, for example, is highly significant by the sign test, 26 out of the 27 pairs showing higher ratios for NB. (More on this point below.)

3.8. Second, the ratios show some significant variation between passage-pairs within a given set [compare, e.g., sr. nos. 18 and 21]; since the passages are small, this phenomenon should indeed be expected. Even so, the variation is marked only within the second and the third sets. The variation between whole sets of passages is clearly larger and more significant.

3.9. The major finding seems to be that the ratio is of the order of 84% for passages from 'Sheser Kavita', about 104% for passages set in SF Examination for translation into Bengali, about 99% for passages from Nehru's "Autobiography", and nearly 106% for the set of five newspapsages. But except for the "Sheser Kavita" set, no other set of passages seems to show an overall ratio definitely different from 100%.

3.10. To see this, consider, for example, the ratios in the row "sub-total : 7-16" in Table 2. The subject-wise average ratios

vary from 95.9% to 113.0%. Even ignoring that such averages themselves have some margin of uncertainty, being based on 10 sample passage-pairs each written only once, the between subjects variation is too large for the overall ratio 103.6% to be regarded as significantly higher than 100%. The five subjects may be regarded as a random sample from a larger population, and we may be ultimately interested in the average of ratios over this population of subjects. Assuming that the subjectwise averages are normally distributed around the population, the five sample ratios give an estimate of roughly 8% of the standard deviation in the population, so that the standard error of the observed ratio (103.6%) would come to more than 3%. (The same conclusion can be reached by noting that 2 out of the 5 ratios are below 100%)

3.11. The overall ratio (99.1%) for passages from Nehru's "Autobiography" cannot obviously be regarded as significantly different from 100%, but for the set of news-passages, probably, the overall ratio may be taken as significantly above 100%. In this latter case, the true ratio may lie anywhere between 100% and 110%, roughly speaking. Against all these, the set of passages from "Sheser Kawita" shows that the Bengali: English ratio is significantly and appreciably below 100%, of the order of 85%.

3.12. The following straight lines have been passed through Figs. 1 to 4 respectively in addition to the equiangular line (y=x) already mentioned : $y = 0.842x$, $y = 1.036x$, $y = 0.991x$, and

$y = 1.063x$. These indicate the overall ratios for passages of different sets. It can be seen from the Figures that these lines pass centrally through the scatters, points falling almost symmetrically on both sides. But as stated above, it is only in Fig. 1, and probably to a small extent in Fig. 4, that the regression may be said to be significantly deviating from the equiangular line $y=x$.

3.13. The peculiar result for the "Sheser Kavita" set makes it difficult to reach any definite overall conclusion. It may be explained by the highly idiomatic character of the passages from "Sheser Kavita". The passages in the other three sets were on political or factual topics and employed ordinary language which is easier to translate from Bengali to English or vice versa. Faithful translations of highly idiomatic Bengali or English may have to be "longer" relatively than faithful translations of journalistic writing or stories. (It may also be that, in general, translations would tend to be a little "longer" than the originals, if they are to bring out the flavour of the original, but this point seems to be less important.) If the experiment included highly idiomatic English passages which can be translated into Bengali only with great difficulty, the overall ratio for such passages might easily have been appreciably larger than 100%. The present author is inclined to hold this point of view, and to conclude that, on the whole, the time requirements are roughly of the same order for writing the same message in English and in Bengali. The time requirements are very nearly of the same order for journalistic

and similar passages where translation is fairly easy and satisfactory; but if one considers translations of highly idiomatic Bengali or English the time requirement may be appreciably higher for the translation - but this phenomenon has been observed only with English translations of highly idiomatic Bengali passages and not with Bengali translations of highly idiomatic English passages.

4.1. Limitations of the experiment: Two limitations have already been apparent, viz., the small coverage of types of passages, and the dependence on only five subjects for the writing experiment. But these do not seem to be really serious. Broad conclusions could be drawn in spite of these. There are certain other limitations which are of more serious consequence.

4.2. First of all, it must be noted that all the five subjects had Bengali as their mother tongue. Englishmen or others suitable for the experiment were not available. Theoretically, it is, of course, doubtful how far Bengalees can be representative subjects for writing experiments on English. Also, conscious or unconscious elements of bias might conceivably influence the comparative time requirements.

4.3. To counteract this, it may be stated that all five subjects were graduates, three with English as one of the subjects at the B.A. level, and the other two in science subjects taught through the medium of English. In all cases, the system of education as well as the nature of subsequent occupation required the writing

of English more frequently than the writing of Bengali. Also, the subjects were aware of the purpose of the experiment, and approached the experiment with a scientific frame of mind. They tried to put in equal efforts when writing in the two languages[1].

4.4. This brings us to the consideration of the two really serious limitations of the present experiment. It is felt that taking a practical view of things, they do not seriously vitiate the conclusions reached in Section 3, but they do make the present experiment "unscientific" in some important respects.

4.5. If the subjects could put in exactly equal effots for writing both passages of a pair, they could have avoided all types of bias, conscious or unconscious. But there is no method of guaranteeing the exact equalisation of efforts, which had to be made by the subjects in a completely subjective manner. This was apparently an insoluble problem. Deliberate bias was, of course, eliminated by these attempts, but unconscious elements of bias could easily persist. This, it must be admitted, is an unscientific feature of the experiment although, in the opinion of the present author, such unconscious types of bias have not seriously vitiated the results of the experiment.

1/ In a small number of cases, it so happened, the subject himself and/or the time-keeper felt that a little more effort had been made on one passage of a pair; and these feelings were also recorded. But such cases were few and were also evenly distributed between the two languages; and so the point may be ignored for all practical purposes.

4.6. There is again, a great deal of vagueness in the question sought to be answered through this experiment. Writing speed depends greatly on the quality of handwriting (legibility etc.) desired, and also on the size of the letters. In order that the results of the experiment may be valid bases of comparisons, each subject should write both the passages of each pair equally well using letters of the same size. This also raises a host of almost insoluble problems, viz., definition and measurement of quality of, and size of letters in, handwriting specimens, ensuring equalisation of quality of, and size of letters in, actual writing of the English and the Bengali passages of each pair, and in cases where this equalisation had not been satisfactory, applying suitable corrections or adjustments for the same.

4.7. No instruction was given to the subjects about equalising the quality of handwriting or the size of letters, excepting that they were asked to write each letter distinctly as far as possible. Some typical specimens of handwriting are given below. They seem to show that whatever be the method of measuring quality of, and size of letters in, handwriting specimens, the two passages of each pair were written in very similar manner[1/]. This was partly the effect of trying to write the two passages with equal efforts and partly of the correlation between English and Bengali handwriting of the same subject.

1/ It is also hoped that the specimens of English handwriting and the speed of writing English passages as apparent from Table 1 - between 25 and 38 words per minute, depending on the subject - will show how far the five Bengalee subjects did justice to the English script.

Farewell, My Friend. Date 29.7.58.

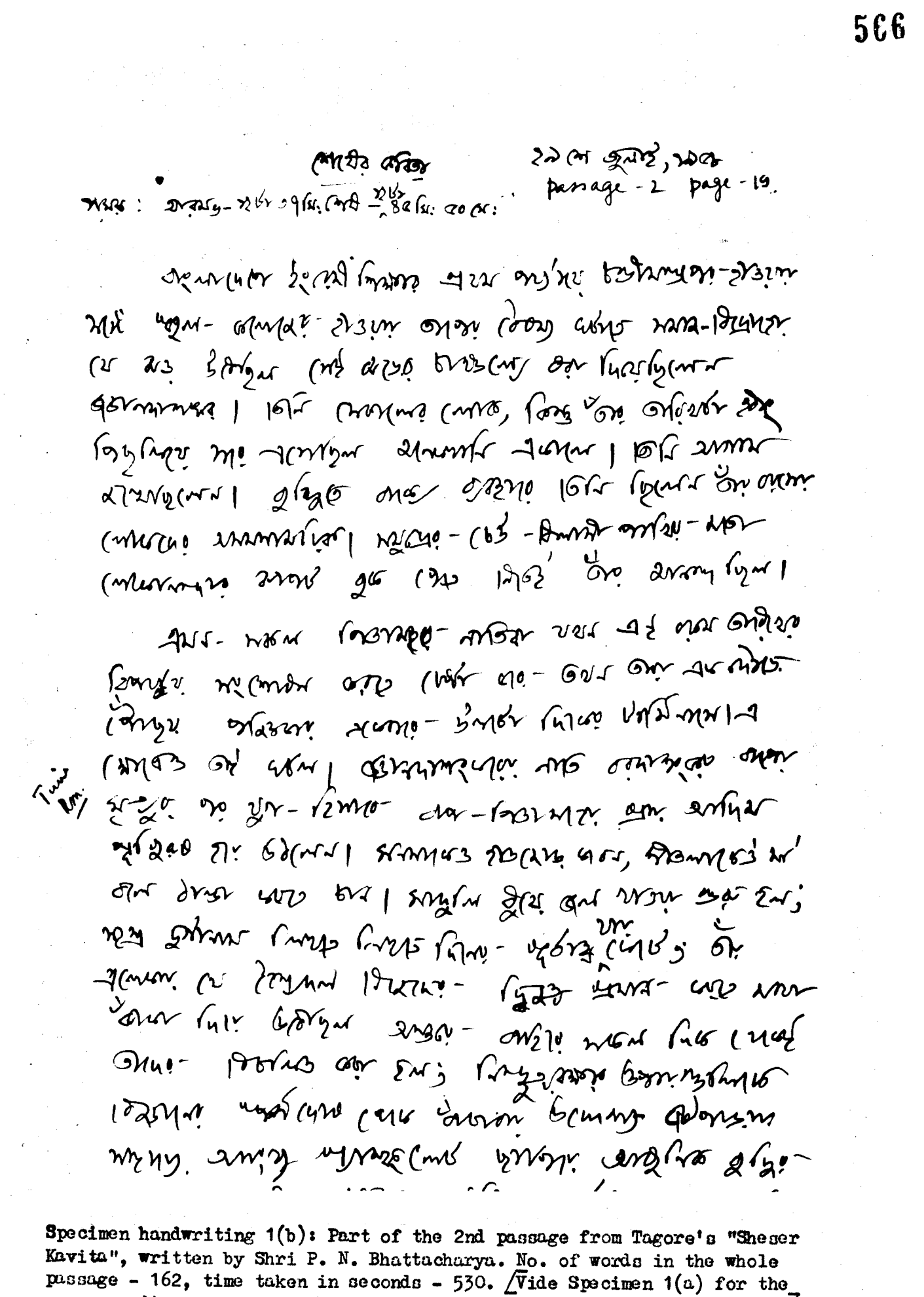
Start.- 4-13-[illegible] secs P.M. - finish - 4-23-2 secs P.M.

Passage - II. Page - 19.

The brisk phase of the introduction of English education into Bengal was marked by a storm of social conflict, generated by the uneven distribution of atmospheric pressure between the old seats of learning and the new schools and Colleges. Jnanadashankar had been caught in this storm. He belonged by birth to the old generation, but had been suddenly blown far ahead into the new. Born before his time, neither in outlook, nor in speech, nor in habits did he contemporise with his contemporaries. Like the sea-bird that revels in rocking on the wave, he loved to bare his breast to the blast of social calumny.

When the progeny of such a grandfather take upon themselves to set right the freaks of the calendar, they usually head straight for the opposite terminus. Which is what happened to Jnanadashanka's grandson, Vanadashanka. This gentleman, after his father's death, contrived to become anachronously the remote ancestor of his immediate progenitors. He was a devotee of serpent-goddess and shunned the morning worship on the one thousand names of the

Specimen handwriting 1(a): Part of the 2nd passage from "Farewell My Friend", [English translation of Tagore's "Sheser Kavita"] written by Shri P. N. Bhattacharya. No. of words in the whole passage - 240, time taken in seconds - 557. [Vide Specimen 1(b) for the corresponding Bengali passage in "Sheser Kavita".]



Specimen handwriting 1(b): Part of the 2nd passage from Tagore's "Sheser Kavita", written by Shri P. N. Bhattacharya. No. of words in the whole passage - 162, time taken in seconds - 530. [Vide Specimen 1(a) for the corresponding passage in the English translation, "Farewell My Friend".]

(255 secs) ✓

10th passage
Page 186

Start – 1-00-00. Finish 1-4-15

Once upon a time there was a hermit with his long beard and matted hair. He lived in a thatched cottage in the forest. He had strange magical powers through which he could perform wonders. People would visit his cottage in large numbers and would get from him whatever they wished for.

There was a mouse in a hole in the hermits cottage. He saw many persons came there every day and received the favour of the hermit. One day the mouse thought he too would have his desires fulfilled; so he walked timidly up to the hermit one morning and bowed low. The hermit asked the mouse what he wanted. Naturally the ambition of the mouse was to be a cat. The hermit muttered something and moved his hand and what wonder – the mouse became transformed into a cat within the twinkling of an eye.

Specimen handwriting 2(a): The 10th of the English passages set in a public examination, written by Shri R. N. Mukherjee. No. of words - 146, time taken in seconds -255. [Vide Specimen 2(b) for the corresponding Bengali passage.]

10th Passage - Pol. 187

Start — 1-08-00 (272 Secs) ✓ Finish — 1-12-32

[illegible]

Specimen handwriting 2(b): The Bengali translation of the 10th English passage set in a public examination, written by Shri R. N. Mukherjee. No. of words - 121, time taken in seconds - 272. [Vide Specimen 2(a) for the corresponding English passage.]

4.8. In conclusion, it may be emphasised that, in the opinion of the present author, the above limitations do not vitiate the broad conclusion given at the end of Section 3.

5.1. Concluding Observations : The current Bengali script is popularly regarded as appreciably less efficient than the English (or roman) script which is apparently more suited for cursive writing. The present experiment contradicts this popular impression by showing that if the same message has to be expressed in writing in English and in Bengali, about the same time would normally be required for writing in these two languages. Spoken Bengali, it may be presumed, is not appreciably more efficient than spoken English.

5.2. The popular impression has its origin in some confusion between letters and syllables. The Bengali script employs the syllabic system inherited from Brāhmi [vide S. K. Chatterjee, 1935]. Each character is, generally speaking, a syllable by itself. This requires the use of conjunct consonants and also of special abbreviated forms of vowels occurring immediately after consonants. As a consequence, the script is more complicated than the roman system which shows each letter clearly and separately in a linear sequence. But while a character in the Bengali script is more complicated and takes more time in writing than a single letter in the roman system, there is no reason why a character in the Bengali script would take more time than the corresponding combination of letters in the roman system. Consider, for example, the syllable 'strai' in

Bengali (written). In the Bengali script, all the sounds are written close together, while the roman system would stretch them into a linear sequence. The Bengali system is, of course, more complicated and is also quite irritating, when one is writing fast, for the hand does not seem to move(!) but in fact it is not much slower than the roman system, as the present experiment shows.

5.3. This inevitably reminds one of Professor S. K. Chatterjee's proposal [Chatterjee, 1935] for introducing a roman type of script (or alphabet) for Bengali and other languages of India. No experimentation could be done to compare the efficiency of the proposed script with that of the current Bengali script. Persons sufficiently habituated to the romanised system of writing were not available for the experiment. Probably the two systems would be more or less equally timeconsuming so far as handwriting is concerned[1]. However, the romanised system possesses many important advantages, not the least of which would be the considerable saving in time and cost for printing.

1/ It may be noted here that a large proportion of 'a'-sounds ('a' as in 'fall') are not shown in the current Bengali script, but would have to be shown in the romanised script.

References[1/]

Aitchison, J. and Brown, J.A.C.(1957) : The Lognormal Distribution, with special reference to its uses in economics. Cambridge University Press.

Baker, S. J. (1950) : The pattern of language. J. Gen. Psychol., Vol. 42, 25-66.

*Bandyopadhyay, Shrikumar (1956) : Bangla Sahitya Upanyaser Dhara, 3rd Edition, Modern Book Agency, Calcutta.

*_____ (1959) : Bangla Sahityer Bikasher Dhara, Orient Book Company, Calcutta.

Barnard, G. A. (1951) : The theory of information. (With discussion) Jour. Roy. Statist. Soc., Series B, Vol. XIII, 46-64.

Bhattacharya, N. (1960) : Syllable counts on modern Bengali prose. Presented to the 47th Session of the Indian Science Congress, Bombay, Abstract in Part III of Proceedings, p. 37.

_____ (1961) : Sampling experiments on the combination of independent χ^2-tests. Sankhyā, Vol. 23, Series A, 191-196.

_____ (1961) : Correlation between word-length and word-frequency. Presented to the 48th Session of the Indian Science Congress, Roorkee, Abstract in Part III of Proceedings, pp. 28-29.

_____ and Mahalanobis, B. (1964) : A concentration curve for regional disparity. In Studies Relating to Planning for National Development, mimeographed, Indian Statistical Institute, Calcutta.

Bhattacharya, N. and Mukherjee, R. N. (1965) : On fitting the Pareto law to income distributions. Arthaniti, Vol. VII, No.II, 177-82.

Bell, D. A. (1953) : Information Theory and Its Engineering Applications. Sir Isaac Pitman & Sons, Ltd., London, Second Edition, 1957.

_____ and Ross, A.S.C. (1956) : Negative entropy of Welsh Words. Pp. 149-153 of Information Theory, 3rd London Symposium, 1955, Ed. by E. Colin Cherry. Butterworths Scientific Publications, London.

1/ The asterisk (*) indicates Bengali works.

Birnbaum, Z. W. (1953) : Distribution-free tests of fit for continuous distribution functions. Ann. Math. Stat., Vol.24, 1-8.

Bodmer, Frederick (1945) : The Loom of Language. George Allen and Unwin Ltd., London.

Booth, A. D., Brandwood, L. and Cleave, J. P. (1958) : Mechanical Resolution of Linguistic Problems. Butterworths Scientific Publications, London.

Bourne, Charles P. and Ford, Donald B. (1961) : A study of the statistics of letters in English words. Information and Control, Vol. 4, 48-67.

Bradley, A. C. (1960) : Shakespearean Tragedy : Lectures on Hamlet, Othello, King Lear, Macbeth. MacMillan & Co. Ltd., London. (First Published November 1904).

Brinegar, Claude S. (1963) : Mark Twain and the Quintus Curtius Snodgrass letters : a statistical test of authorship. Jour. Amer. Stat. Assn., Vol. 58, 85-96.

Burton, N. G. and Licklider, J. C. R. (1955) : Long-range constraints in the statistical structure of printed English. Amer. Jour. Psychol., Vol. 68, 650-653.

Carroll, J. B. (1938) : Diversity of vocabulary and the harmonic series law of word-frequency distribution. Psychol. Rec., Vol. 2, 379-386.

Chapanis, A. (1954) : The reconstruction of abbreviated printed messages. J. Exptl. Psychol., Vol. 48, No. 6,

Chatterjee, Suniti Kumar (1921) : A brief sketch of Bengali phonetics. Reprinted for International Phonetic Association from "Bulletin of the School of Oriental Studies", London, Vol.II,Part 1.

_____ (1926) : The Origin and Development of the Bengali Language. University of Calcutta.

_____ (1935) : A Roman Alphabet for India. Calcutta University Phonetic Studies, No.4, Calcutta Univ. Press.

_____ (1945) : Bhasa Prakash Bangla Byakaran, 3rd Edition, University of Calcutta.

Cherry, E. Colin (1951) : A history of the theory of information. Report of Proceedings of the Symposium on Information Theory, London, 1950, Ministry of Supply, ed. by Willis Jackson, Pp. 22-43, Discussion Pp. 167-168.

Cochran, W. G. (1952) : The χ^2-test of goodness of fit. Ann. Math. Stat., Vol. 23, 315-345.

_____ (1954) : Some methods of strengthening the common χ^2-tests. Biometrics, Vol. 10, 417-451.

_____ (1963) : Sampling Techniques. 2nd Edition. John Wiley & Sons. Inc., New York.

Corbet, A. S., Fisher, R. A. and Williams, C. B. (1943) : The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Eco., Vol. 12, 42-58.

Cox, D. R. and Brandwood, L. (1959) : On a discriminatory problem connected with the works of Plato. Jour. Roy. Stat. Soc., Series B, Vol. 21, 195-200.

*Das, Jnanendra Mohan (1916) : Bangala Bhasar Abhidhan. 2nd Revised Enlarged Edition in 2 Vols, 1937. Indian Publishing House, Calcutta.

Deb Chowdhury, Probodh Chandra (1931) : Word Frequency in Bengali and Its Relation to the Teaching of Reading. Dacca Univ. Bulletin, No. XIV. Published by Univ. of Dacca, Ramna, Dacca.

Dewey, G. (1923) : Relative Frequency of English Speech Sounds. Harvard University Press, Rev. Edition, 1950.

Elderton, W. P. (1945) : Criket scores and some skew correlation distributions (an arithmetical study). Jour. Roy. Stat. Soc., Series A, Vol. 108, 1-11.

_____ (1949) : A few statistics on the length of English words. Jour. Roy. Stat. Soc., Series A, Vol. CXII, 436-445.

Eldridge, R. C. (1911) : Six Thousand Common English Words. Buffalo, Clement Press [Referred to by Zipf (1949) and many others].

Feller, W. (1957) : An Introduction to Probability Theory and Its Applications, Vol.1, 2nd Edition, John Wiley & Sons, Inc., New York.

Flesch, Rudolf (1946) : The Art of Plain Talk. Harper & Brothers Publishers, New York and London.

_____ (1948) : A new readability yardstick. J. Appl. Psychol., Vol. 32, 221-233.

French, N. R., Carter, C. W., Jr., and Koenig, W., Jr. (1930) : The words and sounds of telephone conversations. Bell Syst. Tech. Jour., Vol. 9, 290-324.

Fries, C. C. and Traver, A. A. (1940) : English Word Lists. American Council on Education, Washington, D. C.

Fucks, Wilhelm (1952) : On mathematical analysis of style. Biometrika, Vol. 39, 122-129.

_____ (1954) : On nahordnung and fernordnung in samples of literary texts. Biometrika, Vol. 41, 116-132.

_____ (1955) : Mathematical theory of word-formation. Pp. 154-170 of Information Theory, Third London Symposium, 1955, Edited by E. Colin Cherry. Butterworths Scientific Publications, London.

Gaines, H. F. (1956) : Cryptanalysis. Dover Publications, New York.

Garner, W. R. (1958) : Symmetric uncertainty analysis and its implication for psychology. Psychol. Rev., Vol. 65, 183-196.

Gibson, H. N. (1962) : The Shakespeare Claimants : A critical survey of the Four Principal Theories Concerning the Authorship of the Shakespearean Plays. Methuen & Co. Ltd., London.

Goldman, Stanford (1954) : Information Theory. Prentice-Hall, New York.

Good, I. J. (1953) : The population frequencies of species and the estimation of population parameters. Biometrika, Vol. 40, 237-264.

_____ (1957) : Distribution of word frequencies. Nature, London, Vol. 179, 595.

_____ and Toulmin, G. N. (1956) : The number of new species and the increase in population coverage when a sample is increased. Biometrika, Vol. 43, 45-63.

Goodman, L. A. (1954) : Kolmogorov-Smirnov tests for psychological research. Psychol. Bull., Vol. 51, 160-168.

Goulden, C. H. (1939) : Methods of Statistical Analysis. First Edition. John Wiley & Sons Inc., New York.

Hanley, M. L. (1937) : Word Index to James Joyce's Ulysses. University of Wisconsin, Madison, Wisconsin. (Statistical tabulation by M. Joos.)

Harris, Z. S. (1959) : Distributional Structure in Linguistics Today, p. 62, New York.

Herdan, Gustav (1953) : Language in the light of the theory of information. Part I. Metron, Vol. XVII, Nos. 1-2, 89-125.

_____ (1955) : -do- Part II. Metron, Vol. XVII, Nos.3-4, 93-121.

_____ (1956) : Language as Choice and Chance. P. Noordhoff Ltd., Groningen, Holland.

_____ (1957) : The mathematical relation between the number of diseases and the number of patients in a community. Jour. Roy. Stat. Soc., Series A, Vol. 120, 320-330.

_____ (1958a) : The relation between the dictionary distribution and the occurrence distribution of word-length and its importance for the study of quantitative linguistics. Biometrika, Vol. 45, 222-228.

_____ (1958b) : The mathematical relation between Greenberg's index of linguistic diversity and Yule's characteristic. Biometrika, Vol. 45, 268-270.

_____ (1960) : Type-token Mathematics : A Text-Book of Mathematical Linguistics. Mouton & Co., The Hague.

_____ (1961) : A critical examination of Simon's model of certain distribution functions in linguistics. Applied Statistics, Vol. X, No.2, 65-76.

_____ (1962) : Calculus of Linguistic Observations. Mouton & Co., The Hague.

Hoffman, Calvin (1955): The Man Who was Shakespeare. Julian Messner, Inc., New York.

Hornby, A. S., Gatenby, E. V. and Wakefield, H. (1948) : The Advanced Learner's Dictionary of Current English. Oxford Univ. Press, London.

Jespersen, Otto (1922) : Language, Its Nature, Development and Origin. George Allen and Unwin, Ltd., London, Reprinted 1947.

Jones, Daniel (1963) : Everyman's English Pronouncing Dictionary. 12th edition, Dent, London.

Kaeding, F. W. (1897-8) : Häufigkeitswörterbuch der deutsch Sprache. Steglitz bei Berlin : Selbstverlag. [Referred to in Miller (1951) and Herdan (1956)].

Kendall, M. G. (1961) : Natural laws in the social sciences. Inaugural address of the President. Jour. Roy. Stat. Soc., Series A, Vol. 124, 1-19.

_____ (1962) : Rank Correlation Methods. 3rd Edition, Charles Griffin, London.

Kendall, M. G. and Babington Smith, B. (1939) : Tables of Random Sampling Numbers. Tracts for Computers No.24. Cambridge University Press.

Kendall, M. G. and Stuart, A. (1958) : Advanced Theory of Statistics, Vol. 1 — Distribution Theory. Charles Griffin, London.

_____ (1961) : Advanced Theory of Statistics, Vol.2 — Inference and Relationship. Charles Griffin, London.

Khinchin, A. I. (1957) : Mathematical Foundations of Information Theory. Two papers by Khinchin dated 1953 and 1956, Translated by R. A. Silverman and M. D. Friedman. Dover Publications Inc., New York.

Lahiri, D. B. (1951) : A method of sample selection providing unbiased ratio estimates. Inter. Stat. Inst. Bull., Vol. 33(2), 133-140.

_____ (1954) : Technical Paper on Some Aspects of the Development of the Sample Design. National Sample Survey Report No. 5, Ministry of Finance, Government of India.

_____ (1958) : Observations on the use of interpenetrating samples in India. Bull. Inter. Stat. Inst., Vol. 36(3), 144-152.

_____ and Ganguly, A. (1951) : An overall measure of precision of a sample table with applications in the study of relative efficiencies of different sampling units in population censuses. Bull. Inter. Stat. Inst., Vol. 33(4), 55-74.

Linder, Arthur (1963) : Convocation Address at Indian Statistical Institute, 14 April 1963.

Linfoot, E. H. (1957) : An informational measure of correlation. Information and Control, Vol. 1, 85-89.

Lord, R. D. (1958) : Studies in the history of probability and statistics, VIII: De Morgan and the statistical study of literary style. Biometrika, Vol. 45, 282.

Lorge, I. (1937) : The English semantic count.' Teach. Coll. Rec., Vol. 39, 65-77.

Lutoslawski, W. (1897) : The Origin and Growth of Plato's Logic, with an Account of Plato's Style and of the Chronology of his Writings. Longmans, Green & Co., London.

Mahalanobis, P. C. (1946) : Recent experiments in statistical sampling in Indian Statistical Institute.(With discussion.) J. Roy. Stat. Soc., Vol. 109, 325-378.

_____ (1958) : Method of fractile graphical analysis with some surmises of results. Trans. Bose Res. Inst., Calcutta, [illegible], 223-230.

_____ (1960) : A method of fractile graphical analysis. Econometrica, Vol. 28, 325-51; reprinted in Sankhyā, Series A, Vol. 23, 1961, 41-64.

Mandelbrot, Benoit (1953) : An informational theory of the statistical structure of language. Pp. 486-502 of Communication Theory (Proceedings of a Symposium on Communication Theory, London, 1952) Ed. by Willis Jackson. Butterworths Scientific Publications, London.

_____ (1954) : On recurrent noise limiting coding. Pp. 205-221 of Proceedings of Symposium on Information Networks, 1954, Vol. 3, Polytechnic Institute of Brooklyn, New York.

_____ (1959) : A note on a class of skew distribution functions : Analysis and critique of a paper by H. Simon. Information and Control, Vol. 2, 90-99.

_____ (1961a) : Final note on a class of skew distribution functions : Analysis and critique of a model due to H. A. Simon. Information and Control, Vol. 4, 198-216.

_____ (1961b) : Post Scriptum to "Final Note". Information & Control, Vol. 4, 300-304.

Massey, F. J.,Jr. (1951) : The Kolmogorov-Smirnov test for goodness of fit. Jour. Amer. Stat. Assn., Vol. 46, 68-78.

Mcgill, W. (1954) : Multivariate information transmission. Psychometrika, Vol. 19, 97-116.

McMillan, B. (1953) : The basic theorems of information theory. Ann. Math. Stat., Vol. 24, 196-219.

Mendenhall, T. C. (1887)[1/] : The characteristic curve of composition. Science, Vol. IX, March 11, 237-249.

_____ (1901)[1/] : A mechanical solution of a literary problem. The Popular Science Monthly, Vol. LX, December, 97-105.

Miller, G. A. (1951) : Language and Communication. McGraw-Hills Book Co., Inc., New York.

_____ (1957) : Some effects of intermittent silence. Amer. Jour. Psychol., Vol. 70, 311-313.

_____ (1964) : Mathematics and Psychology. London, Wiley.

Miller, G. A. and Friedman, Elizabeth, A. (1957) : The reconstruction of mutilated English texts. Information and Control, Vol. 1, 38-55.

Miller, G. A. and Newman, E. B. (1958) : Tests of a statistical explanation of the rank-frequency relation for words in written English. Amer. Jour. Psychol., Vol. 71, 209-218.

Miller, G. A., Newman, E. B. and Friedman, E. A. (1958) : Length-frequency statistics for written English. Information and Control, Vol. 1, 370-389.

1/ Reviewed by Williams (1956); vide also Gibson (1962), Brinegar (1963).

Murthy, M. N. and Nanjamma, N. S. (1959) : Almost unbiased ratio estimates based on interpenetrating subsample estimates. Sankhyā, Vol. 21, [illegible]-392.

Newman, E. B. and G[illegible]. L. J. (1952) : A new method for analysing printed [illegible]. Jour. Exptl. Psychol., Vol. 44, 114-125.

Newman, Edwin B. and Waugh, Nancy C. (1960) : The redundancy of texts in three languages. Information and Control, Vol. 3, 141-153.

Noether, G. E. (1963) : Note on the Kolmogorov statistic in the discrete case. Metrika, Vol. 7, No.2, 115-

Oettinger, A. G. (1954) : The distribution of word-length in technical Russian. Mechanical Translation, Vol. 1, 38-40.

Ogden, C. K. (1934) : The System of Basic English. Harcourt, Brace, New York.

Pratt, Fletcher (1942) : Secret and Urgent. Blue Ribbon Books, Garden City, New York. [Referred to by Goldman (1954) and others]

Rao, S. Subba (1960) : A study into the sentence-length as a statistical characteristic determining the prose style of an author. The Half-Yearly Journal of the Mysore University, New Series, Section A — Arts, Vol. XX, No. 1, 1-12.

Ross, A. S. C. (1950) : Philological probability problems. Jour. Roy. Statist. Soc., Series B, Vol. XII, 19-59 (with discussion).

Savage, I. R. (1952) : On the independence of tests of randomness and other hypotheses. Jour. Amer. Stat. Assn., Vol. 52, 53-57.

Scheffè, Henry (1943) : Statistical inference in the non-parametric case. Ann. Math. Stat., Vol. 14, 305-332.

Shannon, C. E. (1948) : A mathematical theory of communication. Bell. Syst. Tech. J., Vol. 27, 379-423, 623-656.

_____ (1951) : Prediction and entropy of printed English. Bell. Syst. Tech. J., Vol. 30, 50-64.

_____ and Weaver, Warren (1949) : The Mathematical Theory of Communication. Univ. of Illinois Press. Urbana.

Siegel, Sidney (1956) : Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Co., Inc., New York.

Simon, Herbert A. (1955) : On a class of skew distribution functions. Biometrika, Vol. 42, 425-440.

_____ (1960) : Some further notes on a class of skew distribution functions. Information and Control, Vol. 3, 80-88.

_____ (1961a) : Reply to "Final Note" by Benoit Mandelbrot. Information and Control, Vol. 4, 217-223.

_____ (1961b) : Reply to Dr. Mandelbrot's Post Scriptum. Information and Control, Vol. 4, 305-308.

Siromoney, Gift (1963) : Entropy of Tamil Prose. Information and Control, Vol. 6, No.3, 297-300.

Stuart, A. (1954) : Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. Jour. Amer. Stat. Assn., Vol. 49, 147-157.

_____(1956) : The efficiencies of tests of randomness against normal regression. Jour. Amer. Stat. Assn, Vol. 51, 285-287.

Sukhatme, P. V. (1954) : Sampling Theory of Surveys with Applications. Indian Society of Agricultural Statistics, New Delhi and Iowa State College Press, Ames, Iowa, USA.

Taraporewala, I. J. S. (1951) : Elements of the Science of Language. 2nd Rev. Enl. Edition, Univ. of Calcutta.

Thorndike, E. L. and Lorge, I. (1944) : The Teacher's Word-book of 30,000 Words. Bureau of Publications, Teacher's College, Columbia University.

Wake, W. C. (1950) : Discussion on "Philological probability problems" by A. S. C. Ross. Jour. Roy. Statist. Soc., Series B, Vol. XII, 19-59.

Wald, A. and Wolfowitz, J. (1940) : On a test whether two samples are from the same population. Ann. Math. Stat., Vol. 11, 147-162.

_____(1943) : An exact test for randomness in the nonparametric case based on serial correlation. Ann. Math. Stat., Vol. 14, 378-388.

Wallis, W. A. and Moore, G. H. (1941) : A significance test for time series analysis. Jour. Amer. Stat. Assn., Vol. 36, 401-409.

Walsh, John E. (1962) : Handbook of Nonparametric Statistics. Investigation of randomness, moments etc. D. Van Nostrand Co., Inc., Princeton, New Jersey.

Whitney, William Dwight (1923) : Sanskrit Grammar, Including both the Classical Language and the Older Dialects, of Veda, and Brahmana. 2nd Edition, Harvard University Press.

Whittaker, E. T. and Watson, G. N. (1958) : Course of Modern Analysis. 4th Edition, Cambridge University Press.

Williams, C. B. (1940) : A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika, Vol. 31, 356-361.

_____ (1946) : Yule's "Characteristic" and the "Index of Diversity". Nature, Vol. 157, 482.

_____ (1952) : Statistics as an aid to literary studies. Penguin Science News, No. 24, 99-106.

_____ (1956) : Studies in the history of probability and statistics, IV: A note on an early statistical study of literary style. Biometrika, Vol. 43, 248-256.

Yardi, M. R. (1946) : A statistical approach to the problem of chronology of Shakespeare's plays. Sankhyā, Vol. 7, 263-268.

Yngve, Victor H. (1956) : Gap analysis and syntax. Pp. 106-112 of IRE Transactions of Information Theory, Vol. IT-2, No.3, 1956 Sumposium on Information Theory. Published by Institute of Radio Engineers, Inc., New York.

Yule, G. Udny (1924) : A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. Phil. Trans. Roy. Soc., London, Series B, Vol. 213, p. 21.

_____ (1938) : On sentence-length as a statistical characteristic of style in prose : With applications to two cases of disputed authorship. Biometrika, Vol. 30, 363-390.

_____ (1944) : The Statistical Study of Literary Vocabulary. Cambridge University Press.

Zipf, George Kingsley (1945) : The repetition of words, time-perspective and semantic balance. Jour. Gen. Psychol., Vol. 32, 127-148.

_____ (1949) : Human Behaviour and the Principle of Least Effort : An Introduction to Human Ecology. Addison-Wesley Press, Inc., Cambridge 42, Massachusetts.

.....